




Claire Dupas  
Philippe Houdy  
Marcel Lahmani  
Editors

# Nanoscience

Nanotechnologies  
and Nanophysics

 Springer

  
EUROPEAN MATERIALS  
RESEARCH SOCIETY

Nanoscience

C. Dupas P. Houdy M. Lahmani  
(Eds.)

---

# Nanoscience

Nanotechnologies and Nanophysics

With 502 Figures (8 in color) and 25 Tables



**Claire Dupas, PhD**

Ecole Normale Supérieure de Cachan  
Avenue du Président Wilson, 94235 Cachan Cédex, France  
E-mail: [claire.dupas@dir.ens-cachan.fr](mailto:claire.dupas@dir.ens-cachan.fr)

**Philippe Houdy, PhD**

Université d'Évry  
Boulevard François Mitterrand, 91025 Évry Cédex, France  
E-mail: [philippe.houdy@univ-evry.fr](mailto:philippe.houdy@univ-evry.fr)

**Marcel Lahmani, PhD**

Club Nano-Micro-Technologie de Paris  
Boulevard François Mitterrand, 91025 Évry Cédex, France  
E-mail: [lahmani@univ-evry.fr](mailto:lahmani@univ-evry.fr)

Translation from the French language edition of  
"Les Nanosciences 1 – Nanotechnologies et nanophysique"  
© 2004 Editions Belin, France

Library of Congress Control Number: 2006929446

ISBN-10 3-540-28616-0 Springer Berlin Heidelberg New York  
ISBN-13 978-3-540-28616-5 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media.

[springer.com](http://springer.com)

© Springer-Verlag Berlin Heidelberg 2007

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Data prepared S. Lyle and by SPi using a Springer TeX macro package

Cover design: *design & production* GmbH, Heidelberg, using a figure from the Hanbücken-Neddermeyer collaboration, Appl. Surf. Sci. 234, 307 (2004)

Printed on acid-free paper SPIN 11415800 57/3100/SPi 5 4 3 2 1 0

---

## Preface

This book is an overview of the fundamentals and applications of nanosciences and nanotechnologies, focusing mainly on nanophysics. Our aim as editors is to use a pedagogical approach in the readers' own language in order to provide a grounding in all the major theoretical and experimental aspects of this new generation of science for students preparing a Masters or a PhD, researchers and university professors.

We wish to extend our thanks to Paul Siffert of the European Materials Research Society for his support of this English version and to Stephen Lyle for his excellent translation from the French.

Paris,  
July 2006

*Claire Dupas*  
*Philippe Houdy*  
*Marcel Lahmani*

---

## Foreword to the French Edition

Research scientists have not only a three-fold duty to society, but also a three-fold ambition: to discover, to invent, and to inform. A piece of knowledge that remained forever the exclusive possession of those who built it would, like a hidden treasure, be worthless to all, even those who concealed it.

Scientists and technicians have a duty to share their knowledge. This is precisely the intention of those who have written this book, explaining the achievements of nanoscience and the prospects it offers us.

The book is aimed at engineers, undergraduates and postgraduates alike. Indeed, it should be accessible to anyone with a scientific or technical background. The subject is an exciting one that has blossomed with astonishing rapidity over the last few years. Nanoscience, nanotechnology and nanomaterials have become a central field of scientific and technical activity. Investigations involve state-of-the-art physics and the full force of our present understanding of matter. Indeed, quantum phenomena are omnipresent. Industrial applications are prolific and wide-ranging, including electronics, communications, magnetism, mechanics, and new materials, not to mention biology, to which a whole volume could easily be dedicated.

The first edition was published in French by unanimous decision of the authors, and they are to be congratulated. This will in no way hinder its dissemination in other languages. The contents are clear, to the point, and appealing. I wish the reader the same pleasure and interest I have found in them myself.

There is a great need for good scientific writing today, and here is a perfect example.

Academy of Science, Paris  
December 2004

*Hubert Curien*

---

## Preface to the French Edition

Convinced thou must confess such things there are  
As have no parts, the minimums of nature.

Lucretius, *De Natura Rerum*

Several centuries before the advent of Christianity, nature had already been broken down into atoms. As the culmination of a long line of Greek philosophers, including Democritus and Epicurus, it was Lucretius who formulated this first detailed ‘atomistic’ description of nature in the first century BC. This was not yet nanoscience, but simply science. However, it was not until the beginning of the nineteenth century, some 20 centuries later, that the atomic theory would be scientifically established by the chemists Dalton, Lavoisier, Gay-Lussac, and others. The science of the microscopic would then work its way up through the ranks during the nineteenth and especially the twentieth century. Physics and chemistry as we know them are largely built upon our knowledge of matter on the atomic scale.

Continuing this same trend from the beginning of the 1980s, it would seem that several new chapters have been added to the history of science, associated with the prefix ‘nano’. The notions of nanoscience and nanotechnology pop up in every sector of modern knowledge. But was it really necessary to create this new branch when physics, chemistry, and biology have long been dedicated to understanding phenomena on the smallest scales?

In fact, rather than designating a new chapter in the history of science, these terms would be better interpreted as a new approach to those chapters already written, a new way of accounting for the scientific disciplines. Let us explain this observation.

The nanometer has long been defined: it is one billionth<sup>1</sup> of a meter or one thousandth of a micron, of the same order as the distance between two atoms

---

<sup>1</sup> In this book, one billion will be taken as  $10^9$ .

in a solid (several tenths of a nanometer). What is new is the ability to manipulate matter on scales ever closer to the nanometer. This new knowhow, this new technology, was naturally given the name of nanotechnology. The fabrication of such small objects opened the way to a new field of scientific investigation. Using novel observational methods developed more or less simultaneously, abstract notions such as the wave function of the electron, the ‘image’ of a single atom, or the presence of just one electron have become commonplace features of everyday experience. This newfound familiarity has indeed stimulated a rush of interest in those sciences that have benefitted from it.

A succession of fruitful exchanges has characterised the interplay between the fabrication of ever smaller and better controlled samples and the understanding of their basic properties. Underlying this interplay is the following question: when exactly can we no longer apply the physics we know, the physics of the macroscopic? The answer has come in the form of a new field of physics: the physics of the mesoscopic, i.e., on intermediate scales, which is of course just one of the many facets of the nanosciences. Precisely fixing the boundary of what one would call nanoscience amounts to detecting that size at which the tiny dimensions of a sample become essential to any explanation of its properties. This is tantamount to defining a frontier, and one whose distance will vary depending upon which property of matter concerns us: electronic states, electron transport, propagation of light, and so on. As a general rule, it is the comparison between the size of the sample and the characteristic lengths of the various phenomena, such as electron mean free path, electron or light wavelength, etc., which constitutes the defining criterion for the frontier of nanoscience. Such a manifold criterion already reflects the many fields affected by nanoscience.

One of the founders in this area, Heinrich Rohrer, has left us with another definition of the nanosciences, centered upon the subjects it studies:<sup>2</sup>

Thus we could call nanoscience the science of dealing with nano-individuals. This applies to measuring, understanding and selectively modifying properties, to manipulating, positioning and machining nano-objects as well as to developing new concepts in treating nano-individuals, especially large numbers of them.

One advantage of this definition is that it brings out the extent to which the methods of fabrication and instruments of observation form the very heart of the subject.

Before going further with this theme, it is perhaps wise to ask whether nanoscience and nanotechnology do not merely constitute a passing phase, an overwhelming fascination due to the present successes of microelectronics, or

---

<sup>2</sup> H. Rohrer, The nanoworld: Chances and challenges. In: Proc. of Intl. Conf. on Nanophase Chemistry, Houston USA (23–24 October 1995). H. Rohrer received the Nobel prize for physics 1986, jointly with G. Binnig, for the invention of the scanning tunneling microscope.



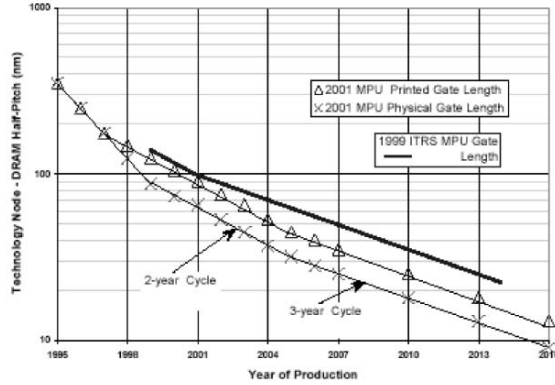
whether they represent a genuine driving force capable of transforming science and technology in some enduring way. The events of the last thirty years have taught us that the techniques of miniaturisation have never ceased to progress over this period. Electronic components are now microcomponents, on their way to becoming nanocomponents. The trend has been continuous and steady, as summed up in the law formulated by Gordon Moore at the beginning of the 1970s. Indeed, it was observed that the number of transistors incorporated in a chip was multiplied by four every three years, and this law has been fulfilled ever since. These developments arise from the increased surface area of chips, which has gradually progressed from  $\text{mm}^2$  to  $\text{cm}^2$ , but also and especially, the reduced size of the components that can now be made. Every generation of integrated circuit is drawn with a design rule (the smallest dimension or minimum feature size) 0.7 times that of the previous generation, whereby all areas are divided by a factor of 2. Successive generations of integrated circuits have come out on a steady 3-year cycle, and recently, at the rate of one every 2 years. The driving force behind this evolution can be found in the simultaneous quest for better performance (reduction of the gate length increases component speed), cost cutting (to manufacture more components from the same slice of silicon), and increased reliability (a single chip can integrate more and more functions). All these features can be summed up in the formula: smaller, faster, cheaper.

But at this rate, where will microelectronics end up? A programme for its future development can be found in *The International Technology Roadmap for Semiconductors*.<sup>3</sup> According to this document, in 2003, the half-pitch of a memory cell array was 100 nm, whilst the gate length of a MOS transistor, printed with a 60-nm pattern, had a physical length of only 45 nm. If the present trend continues, it predicts that 4 generations later in 2015, this physical gate length will be a mere 10 nm (see Fig. 1). The age of nanotechnology is thus well and truly programmed for the beginning of the twenty-first century.

On this scale, all our ideas must be reconsidered, since the approximations to which our observations are referred are no longer valid: atoms can be moved individually, and electrons or photons can be counted one by one. Samples are so small that each must be treated as a quantum object in its own right. Naturally, the operating principles of transistors which have been applied for better or for worse to developments in microelectronics throughout its evolution will now have to be reassessed. Through the very success of their own undertakings, technologists have no choice but to enter uncharted territory, where science alone can guide them.

It is therefore reasonable to ask whether this ‘peaceful’ evolution of the past few decades will continue without upset, or whether the passage into the nanoworld can only happen at the cost of a series of revolutions, each capable

<sup>3</sup> *The International Technology Roadmap for Semiconductors* can be consulted on the Internet: <http://public.itrs.net/Files/2002Update/2002Update.pdf> in portable document format.



**Fig. 1.** Reduction in the gate length of MOSFET transistors as projected in the International Technology Roadmap for Semiconductors. *Horizontal axis:* year of production. *Vertical axis:* minimum feature size (nm). The *thick curve* shows predictions made in 1999 compared with those made in 2001 and confirmed in 2002 (*thin curves*). *Triangles* correspond to printed gate lengths, resulting from the lithographic process, whilst *crosses* correspond to physical gate lengths at the end of the fabrication process

of transforming the basic concepts and technologies, or even the industrial models of microelectronics. Indeed we have already mentioned the quantum revolution. But there is also the revolution of molecular electronics. Instead of the top-down approach, where everything is based upon the power of miniaturisation, will it be possible to substitute a bottom-up approach, capable of assembling mass-produced nanoelements? Will the transistor be replaced by molecules, and copper wire by carbon nanotubes?

These radical transformations, these revolutions, are prepared by scientists in their laboratories. Well ahead of the techniques employed by industry, they are already fabricating much smaller samples than those predicted by the roadmap, isolating molecules, building up elementary components, analysing their properties, and developing the corresponding theory.

The point of all this development is quite clear. Science and technology are more closely bound than ever before, to master such rapidly evolving techniques, understand the properties of novel objects, and predict and demonstrate new principles regarding these components. The field is a large one, as Feynman pointed out in 1959:<sup>4</sup> “There’s plenty of room at the bottom.” Indeed, there is more and more room at the bottom because, since that time, the nanoworld has been extended in the most extraordinary manner. The days

<sup>4</sup> This was the title of a lecture given by R.P. Feynman on 29 December 1959, and available at the website [www.zyvex.com/nanotech/feynman.html](http://www.zyvex.com/nanotech/feynman.html). Feynman was awarded the Nobel prize for physics in 1965.

when the sciences operated in splendid isolation are gone. Chemistry has laid claim to a quite unequalled knowhow in the construction and understanding of molecular and supramolecular objects, whilst biology, which has also reached down to the molecular level, has brought us novel concepts and given access to the great wealth of the living world. Other branches of science, like mechanics, so often forsaken, have proved crucial to the development of microsystems. Every day, the sciences converge a little further at the beckoning of this new technology. And this convergence invites revolution at every level: intellectual, organisational, and educational.

Young scientists today are likely to be attracted by the depth and novelty of the challenge, setting off on a voyage of discovery in an unknown world, where the ways have not yet been signposted. However, university and professional training must prove itself equal to the task, providing them with the theoretical and conceptual background they will require. One of the aims of this book is precisely to provide a tool that can be used to train, not only students, but teachers and research scientists. It has been written by research workers and university teachers who are experts in their own fields and fully up-to-date with the latest developments. It has been put together in such a way as to produce a uniform and complete entity that can be approached directly via any of the chapters, whilst maintaining a high level of complementarity between them.

Chapters 1–6 are concerned with the tools and equipment which make nanoscience possible. Described first are the processes of lithography and engraving, used to fabricate smaller and smaller objects from a bulk material according to what is known as the top-down approach, working from the macroscopic to the nanoscopic. In lithography, the limits of miniaturisation have been gradually pushed back using either UV radiation of ever shorter wavelength, and even X-rays, to reduce diffraction effects, or electron beams and ion beams. This represents a major step forward, that has served as a precondition for the development of miniaturisation.

Self-organisation and self-assembly constitute another approach to fabricating nanometric objects. Here, one must look back to the invention of molecular beam epitaxy (MBE) in the mid-1970s. This provided a simple and reproducible technique for preparing ultrathin films of metals or semiconductors, whose thickness could now be reduced to a single atomic layer, viz., 0.3 nm. The science of objects with one nanometric dimension had begun: quantum wells, superlattices, doping planes reduced to one monolayer would soon be better and better controlled. However, size reduction was pursued still more vigorously in order to benefit fully from the elimination of degrees of freedom in fabricated objects. Naturally occurring physical phenomena had to be put to use to shape matter in this way. Through surface reconstructions and stress relaxation, nanometric objects such as quantum dots, wires, and islands can be prepared directly in an effective manner. By the same methods, a template can be generated for the fabrication of such objects. It is only by techniques of this kind that one can take the bottom-up route to

nanoelectronics, assembling nano-objects directly, as in a child's construction toy.

The appearance of new observation and manipulation tools has also contributed in an essential way to developments in this field. They are based on the use of a very small probe, capable of making measurements locally. The displacement of this probe in the immediate vicinity of the sample using high-precision piezoelectric actuators allows one to follow the variations of the measured parameter and construct a representation of the sample, an image of sorts. What is more, the diversity of these probes has led to a whole family of new instrumentation.

To begin with, there is the scanning tunneling microscope (STM), which registers the electronic current crossing via the quantum tunnel effect from the conducting surface to a very fine metallic point or tip held close by. In fact, this instrument probes the local electron density right down to atomic scales.

The probe can also detect the attraction and repulsion operating between the tip and the surface. In this case, the instrument, known as the atomic force microscope (AFM), is quite capable of observing insulating surfaces. It is used first and foremost to establish the topography of the surface. However, the wide range of forces coming into play leads to an equally wide range of measuring instruments, depending on whether one is concerned with electronic forces, magnetic forces, and so on. In certain cases the tip can serve both as an observation instrument and as a tool for modifying the surface, moving atoms around in a controlled way or activating localised chemical reactions.

In the same category are the scanning near-field optical microscopes (SNOM). In this case the probe is a very fine optical fibre with diameter well below the wavelength of the light it collects and transports to the detector.

This part of the book ends with a description of alternative lithographic methods currently under development, whether they are derived from near-field microscopy in which the probe becomes a writing implement, or from nanoimprint techniques.

Chapters 7–10 describe the various families of nano-objects. Their great diversity and the multitude of methods for preparing them attest to the astonishing vitality of this field of activity. Their unique properties often mean that they have become basic building blocks when assembling complex structures according to the bottom-up approach.

Clusters are atomic assemblies prepared by methods arising in atomic physics. The theoretical basis for their properties, methods of preparation and potential applications are presented in full.

Fullerenes and carbon nanotubes are the best known of the assemblies involving carbon. These newly discovered forms add to the list that began with the long-known graphite and diamond. They are made by rolling up a single sheet of graphite in various ways, some more complex than others. Their theoretical properties and methods of preparation are described in detail. They possess a wealth of properties, leading to a great many applications,

from nanoelectronics to hydrogen storage, not to mention electromechanical nanosystems.

Semiconducting nanowires can also be prepared by direct synthetic methods. As our control over them progresses, they can be used to reproduce most of the functionalities of semiconductors on the nanometric scale. They may well provide original solutions to the requirements of nanoelectronics.

Finally, supramolecular chemistry is the science of self-assembly and self-organisation on the molecular scale. With the great variety of chemical bonds available, they lead to an equally wide range of objects. Their properties may be applied to everything from nanomachines to the components of molecular electronics.

Chapters 11–17 describe the properties and applications of nano-objects from nanoelectronics, through nanomagnetism and information storage to optonics.

Nanoelectronics is a key theme in the development of nanoscience and nanotechnology. How will electronics end up after several decades of miniaturisation? A first stage referred to as ultimate electronics describes the problems and possible solutions that may take current technology to its logical conclusion, with MOSFET gate lengths of 25 nm or less.

Beyond this, components will have to be based on new operating principles. The main solutions currently under investigation are described. The first of these relies on the single-electron transistor, wherein electrons are transferred in a controlled way between conducting islands. Another uses the transfer of magnetic flux quanta between superconducting islands: this is the so-called rapid single-flux quantum (RSFQ). A third solution is based on the properties of small structures, referred to as mesoscopic, across which the electron wave function remains coherent, thereby completely modifying electron transport conditions.

On the other hand, the future may lie in molecular electronics. Could a single molecule be made to behave as a transistor or even as a complex logic function? This is an idea that some have long been reflecting upon and it has stimulated a great deal of research, as much on the theoretical front as in molecular synthesis, or in the measurement of the electronic properties of molecules. The fabrication of components from carbon nanotubes, which have already shown their suitability for achieving complex functions, should also be included in the general area of molecular electronics.

One problem dominates all research in nanoelectronics: this is the need to invent a new architecture for logic circuits, capable of managing the infinite complexity of this electronics.

Nanomagnetism is a key area of research, for magnetic materials are widely used to store information. As the size of memory cells is reduced, a limiting dimension is reached below which the stability of the direction of magnetisation can no longer be guaranteed. How can this limit be pushed back? How can read/write operations be simplified? The discovery of giant magnetoresistance and tunnel magnetoresistance have inspired novel solutions to these problems,

which now underly a new discipline known as spin electronics or spintronics. Electron spin and its interaction with magnetic materials will play ever more crucial roles.

Information storage is one of the basic elements of data processing. This may be achieved by mass memory, or by dynamic memory in constant dialogue with the processor. The various competing technologies are compared and the possibilities for development are discussed.

Optronics is the fourth major field of application of the nanosciences. Indeed, light-emitting components from light-emitting diodes to lasers have fundamentally different properties as their dimensions are reduced. When the size of the structure nears the wavelength associated with electrons and holes, i.e., between 30 and 70 nm depending on the material, the energy levels are shifted and all emission and absorption properties are modified. Because of this, the integration of quantum dots into photonic components has become common practice. The propagation of light itself is significantly affected if the index of the medium is modulated with a period close to the wavelength. The formation of allowed and forbidden bands then transforms the dielectric into a photonic band gap (PBG) material. This new field of investigation has brought optics back into the fold where technology is concerned. Finally, even metallic materials are capable of radiating energy via surface plasmons when in the form of nanostructures. This effect has some rather surprising applications such as the optical sieve effect.

The interface with the biological sciences constitutes the fifth and last field of applications presented. As biological media are by their very essence made up of nanometric units or assemblies of such units within cells, the appearance of new means of investigation or manipulation appropriate to this length scale can be rich in consequences. The emphasis is placed on nanophotonics, presenting the main physical effects: one- or multi-photon fluorescence, spectral broadening, harmonic generation, fluorescence resonance energy transfer (FRET) between particles, near field and evanescent waves, and plasmon resonance already mentioned above.

The various fluorescent molecules and their associations are then discussed, together with assembly techniques used to implement them: coupling of biomolecules and nanoparticles, genetic engineering, and grafting of fluorescent molecules.

Specialised equipment has been designed to implement these techniques: fluorescence microscopes, single-molecule spectroscopy, optical tweezers, non-linear microscopy (two- or three-photon, second and third harmonic, coherent anti-Stokes Raman scattering), near-field optics, fluorescence correlation and photon emission statistics.

Applications of nanophotonics to the study of biological phenomena are then presented through a selection of examples: cell media (membrane openings, dynamics of membrane diffusion, intracellular transport) and biomimetic systems (artificial membranes, Langmuir–Blodgett layers and aqueous gels,

application of the FRET method to analyse the effect of calcium ions on protein folding).

Chapter 18 deals with modelling and simulation. It begins by discussing the relevance of simulation as a theoretical tool in atomic-scale microscopy, then describes the main approaches to simulation. Semi-empirical or empirical simulation, in which the quality depends on the accuracy of the chosen interatomic potential, are covered first. These methods can be improved by taking quantum effects into account. Finally, *ab initio* simulations start from a description of the particle ensemble that is as complete as possible. The complexity of the problem and power of our computers nevertheless impose limits on the size of the objects that can be modelled in this way, making various approximations unavoidable, if computation times are to be kept at a reasonable level. The various limitations of these methods are also described.

CEA, Grenoble, December 2004

*Jean-Louis Pautrat*

### **Acknowledgements**

We would like to thank all members of the French nanoscience community (CNRS, CEA, universities, Grandes Ecoles, industry) who gave a very favourable welcome to the writing of these pedagogical introductions to nanophysics, nanotechnology and nanomaterials, without which they would have been impossible. Special thanks go to those who contributed to the books.

We would also like to thank Hubert Curien of the Academy of Sciences (Paris), who agreed to write the foreword, and Patrice Hesto who gave invaluable advice when the project first began.

We warmly acknowledge the material and financial support of the French Ministry of Research, orchestrated by Jean-Louis Robert of the Department of Physics, Chemistry, and Engineering Sciences, and Michel Lanoo, Director of the Department of Physical Sciences and Mathematics at the CNRS.

Likewise, our warmest thanks go to Claude Puech, President of the Club NanoMicroTechnologie, everyone at the LMN (Laboratoire d'étude des Milieux Nanométriques at the University of Evry, France) and the GIFO (Groupement des Industries Françaises de l'Optique) for their administrative and logistical support.

Finally, we would like to thank Margrit Hanbücken and Alain Perez, Claude Bocarra and Claude Chappert for their continued scientific support, especially during copy-editing sessions, and Paul Siffert of the European Materials Research Society for supporting the English edition of the book.

*Marcel Lahmani, Claire Dupas and Philippe Houidy*

---

# Contents

---

## Part I Tools for Nanoscience

---

### 1 Lithography and Etching Processes

<i>D. Mailly, C. Vieu</i> .....	3
1.1 Definitions and General Considerations .....	3
1.2 Photoresists .....	3
1.2.1 Example of Processing with a Polymer Resist .....	4
1.2.2 Sensitivity and Contrast .....	5
1.2.3 Example of a Positive Resist .....	7
1.2.4 Transfer Stage .....	8
1.3 Subtractive Pattern Transfer .....	9
1.3.1 Wet Etching .....	9
1.3.2 Dry Etching .....	11
1.3.3 Reactive Ion Etching .....	13
1.4 Additive Pattern Transfer .....	15
1.4.1 Lift-Off .....	15
1.4.2 Electrolytic Growth .....	16
1.5 Lithography .....	19
1.5.1 Overview of Lithographic Methods .....	19
1.5.2 Proximity and Contact Photolithography .....	20
1.5.3 Projection Photolithography .....	23
1.5.4 X-Ray Photolithography .....	26
1.5.5 Extreme UV Lithography .....	28
1.5.6 Electron Projection Lithography .....	28
1.5.7 Ion Projection Lithography .....	29
1.5.8 Electron Beam Lithography .....	30
1.5.9 Focussed Ion Beam (FIB) Lithography .....	35
1.5.10 Conclusion .....	39
References .....	40



## 2 Growth of Organised Nano-Objects on Prepatterned Surfaces

<i>M. Hanbücken, J. Eymery, S. Rousset</i> .....	41
2.1 Physical Phenomena in Substrate Pre patterning and Periodic Growth of Adsorbates .....	42
2.1.1 Surface Crystallography: Surface Energy and Surface Stress .....	42
2.1.2 Self-Organised Surfaces: Discontinuities in the Surface Stress .....	47
2.1.3 3D Growth: Energy Criterion and Competition Between Bulk Elastic Energy and Surface Energy.....	48
2.1.4 Role of the Chemical Potential as Driving Force Behind Adsorbate Growth. Curvature Effect and Elastic Stresses.....	52
2.2 Physical and Chemical Methods for Producing Nano-Objects .....	53
2.3 Growth of Nano-Objects on a Naturally Prepatterned Surface Using Its Intrinsic Properties .....	56
2.3.1 Growth of Self-Organised Surfaces .....	56
2.3.2 Uses for Growth on Vicinal Surfaces .....	57
2.4 Growth of Quantum Dots on a Prepatterned Surface by Imposing a Controlled Artificial Pattern .....	58
2.5 Growth of Nano-Objects on a Prepatterned Vicinal Surface by Combining Natural and Artificial Patterning.....	61
2.5.1 Pre patterning the Si(111) Vicinal Surface.....	61
2.5.2 Growth of Gold Nano-Objects on Prepatterned Si(111)....	63
2.6 Conclusion .....	64
References .....	65

## 3 Scanning Tunneling Microscopy

<i>D. Stiévenard</i> .....	69
3.1 Introduction.....	69
3.1.1 General Principles .....	69
3.1.2 General Setup .....	70
3.1.3 Tip Preparation .....	71
3.2 Tunnel Current .....	72
3.2.1 Tunnel Effect Between Tip and Sample .....	72
3.2.2 Tunnel Current: Tersoff–Hamann Theory .....	73
3.2.3 Extending the Tersoff–Hamann Theory .....	73
3.2.4 Resolution .....	74
3.2.5 Contrast .....	75
3.2.6 Measuring the Barrier Height .....	76
3.2.7 Examples .....	77
3.3 STM Spectroscopy .....	80
3.3.1 Elastic Current.....	80
3.3.2 Measuring the Band Gaps of III–V Semiconductors .....	82
3.3.3 Spectroscopy of Individual Quantum Dots .....	82
3.3.4 Inelastic Tunnel Current .....	84

3.4	Tip-Sample Interaction . . . . .	85
3.4.1	Manipulation Modes . . . . .	85
3.4.2	Local Chemistry . . . . .	87
3.5	Conclusion . . . . .	88
	References . . . . .	89
<b>4 Atomic Force Microscopy</b>		
	<i>C. Frétiigny</i> . . . . .	91
4.1	The Device . . . . .	91
4.2	The Various Imaging Modes . . . . .	92
4.3	Image Resolution . . . . .	95
4.4	Contact Mode: Topography, Elasticity and Adhesion Imaging . . . . .	98
4.4.1	Friction Mode . . . . .	100
4.5	Resonant Modes . . . . .	101
4.5.1	General Principles . . . . .	101
4.5.2	Linear Resonant Mode . . . . .	102
4.5.3	Nonlinear Resonant (Tapping) Mode . . . . .	103
4.6	Force Measurements . . . . .	107
4.6.1	Non-Contact Measurements . . . . .	108
4.6.2	Elasticity and Adhesion Measurements on a Single Molecule . . . . .	108
4.7	Magnetic and Electrical Measurements . . . . .	109
4.7.1	Magnetic Measurements . . . . .	109
4.7.2	Electrical Measurements . . . . .	109
4.8	Measuring Mechanical Properties . . . . .	112
4.8.1	Nanoindentation . . . . .	112
4.8.2	Measuring Contact Stiffness . . . . .	112
4.8.3	Contact Resonance Frequency . . . . .	113
4.8.4	Friction Forces . . . . .	114
4.9	Applications in Nanotechnology . . . . .	115
4.10	Conclusion . . . . .	118
	References . . . . .	118
<b>5 Near-Field Optics: From Experiment to Theory</b>		
	<i>C. Boccara, R. Carminati</i> . . . . .	121
5.1	Basic Ideas and the Nature of the Problem . . . . .	121
5.1.1	Resolution, Near Field and Far Field . . . . .	121
5.1.2	Brief History of Near-Field Methods . . . . .	122
5.1.3	Near-Field Optical Microscopy: For What Purpose? . . . . .	123
5.2	Photon Scanning Tunneling Microscope (PSTM) . . . . .	123
5.2.1	Frustration of Evanescent Fields . . . . .	124
5.2.2	PSTM Probe in an Evanescent Field: Scattering Model . . . . .	124
5.2.3	Applications of PSTM . . . . .	126
5.3	Apertureless Near-Field Microscope . . . . .	126
5.3.1	Nano-Antenna Radiating to the Far Field . . . . .	127
5.3.2	Source of Contrast: Scattering Sphere Model . . . . .	127

5.3.3	Sharp-Point Effect. Tip Resolution and Efficiency	129
5.3.4	Field Enhancement Near a Metal Tip	129
5.3.5	Apertureless SNOM: Typical SNOM Setup	131
5.4	Aperture SNOM	133
5.4.1	Metal-Coated Fibre	133
5.4.2	Energy Transmission in a Tapered Metal-Coated Fibre	134
5.4.3	Applications of Aperture SNOM	136
5.5	Plane Wave Expansion. Diffraction Limit	137
5.5.1	Propagation of a Beam in Vacuum	138
5.5.2	Uncertainty Relations and Diffraction	140
5.5.3	Diffraction Limit	140
5.6	Beyond the Diffraction Limit:	
	Near Field and Evanescent Waves	142
5.6.1	Evanescent Waves. Length Scales	142
5.6.2	Uncertainty Relations Revisited	143
5.7	Electromagnetic Radiation. Near Field and Far Field	143
5.7.1	Radiation from an Elementary Source (Electric Dipole)	143
5.7.2	Far-Field Radiation. Diffraction Limit Revisited	144
5.7.3	Near-Field Radiation. Quasi-Static Limit	145
5.7.4	Towards a Model	146
5.8	Dipole Emission Near a Nanostructure	146
5.8.1	Radiative Damping of Dipole Emission	147
5.8.2	Free-Space Dipole Emission	148
5.8.3	Dipole Emission Near an Object	149
5.8.4	Link with the Quantum Approach	150
5.8.5	A Simple Example: Dipole Emission Near a Plane Mirror	151
5.8.6	Dipole Emission Near a Nanoparticle.	
	Radiative and Non-Radiative Coupling	152
	References	155
<b>6 Emerging Nanolithographic Methods</b>		
	<i>Y. Chen, A. Pépin</i>	157
6.1	Introduction	157
6.2	Nanoimprint Lithography	158
6.3	Applications of Nanoimprint Lithography	162
6.3.1	Microelectronics	162
6.3.2	Nanomagnetism	163
6.3.3	Nano-Optics	164
6.3.4	Chemistry and Biology	165
6.4	UV Nanoimprint Lithography (UV-NIL)	166
6.5	Nanoembossing	167
6.6	Soft Lithography	169
6.7	Near-Field Lithography	172
6.8	Conclusion	174
	References	174

---

**Part II Nanoscale Objects**


---

**7 Clusters and Colloids**

<i>A. Perez, P. Mélinon, J. Lermé, P.-F. Brevet</i> .....	179
7.1 Equilibrium Shape .....	180
7.1.1 Liquid-Drop Model .....	180
7.1.2 Wulff Polyhedron .....	182
7.1.3 Beyond the Wulff Polyhedron .....	183
7.1.4 Van der Waals Binding .....	189
7.1.5 Covalent Binding .....	190
7.1.6 Ionic Binding .....	192
7.2 Characteristic Quantity: Radius .....	193
7.2.1 Thermodynamic Quantities: Melting Temperature .....	193
7.2.2 Electronic Quantities .....	196
7.3 Characteristic Quantity: Fluctuations .....	200
7.3.1 Melting Temperature .....	200
7.3.2 Kubo Model .....	202
7.4 Specific Quantum Effects in Nanoscale Systems and Collective Excitations .....	206
7.4.1 Electronic Shell Structure .....	207
7.4.2 Electronic Supershells .....	217
7.4.3 Optical Properties. Collective Excitations .....	225
7.5 Preparation Methods .....	241
7.5.1 Gas Phase Physical Methods .....	241
7.5.2 Liquid Phase Chemical Methods .....	246
7.6 Cluster or Colloid Assemblies .....	252
7.6.1 Assemblies of Metallic Clusters .....	253
7.6.2 Deposition Techniques for Clusters and Colloids .....	254
7.6.3 Characteristic Mechanisms for the Formation of Nanostructures by Cluster Assembly ..	256
7.6.4 Examples of New Nanostructured Systems Prepared by Cluster Deposition .....	260
7.7 Conclusion and Prospects .....	266
References .....	277

**8 Fullerenes and Carbon Nanotubes**

<i>J.-P. Bourgoin, A. Loiseau, J.-F. Nierengarten</i> .....	279
8.1 Introduction .....	279
8.2 Nanotubes and the Crystalline Forms of Carbon .....	280
8.2.1 Diamond and Graphite .....	280
8.2.2 Discovery of Fullerenes .....	281
8.2.3 Discovery of Carbon Nanotubes .....	281
8.3 Fullerenes .....	282
8.3.1 Structure of Fullerenes .....	282

8.3.2	Production of Fullerenes . . . . .	284
8.3.3	Physicochemical Properties of Buckminsterfullerene . . . . .	285
8.4	Carbon Nanotubes . . . . .	289
8.4.1	Crystal Structure of Nanotubes . . . . .	289
8.4.2	Electronic Structure of Carbon Nanotubes . . . . .	292
8.4.3	Self-Organisation of Nanotubes . . . . .	300
8.4.4	Chemical Varieties of Nanotubes . . . . .	301
8.4.5	Synthesis of Nanotubes . . . . .	302
8.4.6	Growth Mechanism for Carbon Nanotubes . . . . .	305
8.4.7	Observation of Nanotubes . . . . .	308
8.4.8	Properties of Nanotubes . . . . .	311
8.4.9	From Science to Applications . . . . .	313
8.5	Conclusion . . . . .	318
	References . . . . .	321

## 9 Nanowires

	<i>J.-C. Labrune, F. Palmino</i> . . . . .	325
9.1	Fabrication of Nanowires . . . . .	326
9.2	The Top-Down Approach . . . . .	327
9.2.1	Soft Lithography . . . . .	327
9.2.2	Near-Field Lithography . . . . .	328
9.3	The Bottom-Up Approach . . . . .	332
9.3.1	Self-Assembly on a Surface . . . . .	332
9.3.2	VLS Synthesis . . . . .	334
9.3.3	Use of Porous Matrices . . . . .	335
9.4	Electrical Conduction in Nanowires . . . . .	335
9.4.1	Electrical Contacts . . . . .	336
9.4.2	Incoherent Transport . . . . .	342
9.4.3	Atomic Chains and Molecules . . . . .	342
9.5	Conclusion . . . . .	344
	References . . . . .	344

## 10 Nano-Objects

	<i>J.-F. Nierengarten, J.-L. Gallani, N. Solladié</i> . . . . .	349
10.1	Dendrimers . . . . .	349
10.1.1	Divergent Synthesis . . . . .	350
10.1.2	Convergent Synthesis . . . . .	352
10.2	Supramolecules . . . . .	353
10.2.1	Self-Assembly by 3D Template Effect Induced by a Metal Cation . . . . .	354
10.2.2	Self-Assembly by Hydrogen Bonding . . . . .	359
10.2.3	Self-Assembly by Hydrophobic Interactions, $\pi$ -Interactions and Charge Transfer Interactions . . . . .	363
10.2.4	Molecular Machines . . . . .	366

10.3 Polymolecular Assemblies ..... 368  
 10.3.1 Self-Assembly in the Bulk ..... 369  
 10.3.2 Self-Assembly on Surfaces ..... 372  
 References ..... 378

**Part III Properties and Applications**

**11 Ultimate Electronics**

*S. Galdin-Retailleau, A. Bournel, P. Dollfus* ..... 383  
 11.1 CMOS Technology ..... 386  
 11.2 MOSFET Scaling ..... 392  
 11.2.1 Basic Principles ..... 392  
 11.2.2 Short Channel Effects ..... 392  
 11.2.3 Scaling Rules ..... 393  
 11.2.4 State of the Art: ITRS Roadmap ..... 395  
 11.2.5 Interconnects ..... 398  
 11.3 NanoMOS Devices ..... 400  
 11.3.1 Specific Problems ..... 400  
 11.3.2 Alternatives to Conventional MOSFET Devices ..... 408  
 11.4 Conclusion ..... 412  
 References ..... 413

**12 Alternative Electronics**

*J.-N. Patillon, D. Maily* ..... 417  
 12.1 Characteristic Length Scales for Nanoscopic Components ..... 418  
 12.2 Single-Electron Devices (SED) ..... 419  
 12.2.1 Basic Ideas ..... 419  
 12.2.2 Transport by Coulomb Blockade ..... 420  
 12.2.3 Double Tunnel Junction ..... 425  
 12.2.4 Single-Electron Transistor ..... 427  
 12.3 Quantum Interference in Nanostructures ..... 428  
 12.3.1 Introduction ..... 428  
 12.3.2 Conductance and Transmission. The Landauer Formula ... 430  
 12.3.3 Calculating the Correction ..... 433  
 12.3.4 Effect of Magnetic Fields ..... 434  
 12.3.5 Universal Conductance Fluctuations ..... 435  
 12.3.6 Cutoffs ..... 437  
 12.4 An Example of Interference: Aharonov–Bohm Effect ..... 437  
 12.5 Superconducting Nanoelectronics: RSFQ Logic ..... 440  
 12.5.1 Introduction ..... 440  
 12.5.2 Superconducting Logic Components ..... 440  
 12.5.3 Structure and Performance of RSFQ Components ..... 442  
 References ..... 446

**13 Molecular Electronics**

<i>J.P. Bourgoin, D. Vuillaume, M. Goffman, A. Filoramo</i> .....	447
13.1 Basic Building Blocks: Choice, Wealth, Complexity .....	448
13.2 A Little History .....	449
13.3 Molecular Components .....	450
13.3.1 Electrodes and Contacts .....	450
13.3.2 Relationship Between Molecular Structure and Properties ..	456
13.3.3 Functions .....	469
13.4 Components Based on Nanotubes .....	479
13.4.1 Field-Effect Transistors .....	479
13.4.2 Single-Electron Transistors (SET) .....	486
13.5 From Components to Circuits .....	490
13.5.1 Fabrication Techniques .....	490
13.5.2 Circuit Architecture .....	497
13.6 Conclusion .....	498
References .....	499

**14 Nanomagnetism and Spin Electronics**

<i>C. Chappert, A. Barthélémy</i> .....	503
14.1 Nanomagnetism .....	504
14.1.1 Vacuum Magnetostatics .....	504
14.1.2 Magnetism in Matter: Fundamental Relations .....	505
14.1.3 Magnetism in Matter: Continuum Approximation .....	511
14.1.4 Novel Magnetic Effects on the Nanoscale .....	525
14.1.5 Magnetisation Dynamics in Magnetic Nanostructures .....	540
14.2 Spin Electronics .....	552
14.2.1 Description .....	552
14.2.2 Origins and Mechanisms of Spin Electronics .....	559
14.2.3 Magnetoresistance of Tunnel Junctions .....	568
References .....	578

**15 Information Storage**

<i>D. Fraboulet, Y. Samson</i> .....	583
15.1 Mass Memories .....	583
15.1.1 Mass Memories: The Hard Disk .....	585
15.1.2 Beyond the Hard Disk. Local Probe Techniques .....	593
15.2 Matrix Memories .....	595
15.2.1 General Principles of Matrix Storage .....	595
15.2.2 Difficulties in Reducing Memory Cells to Nanoscale Sizes ..	599
15.2.3 Matrix Memory Technology in Current Use .....	600
15.2.4 Memory Concepts Under Development .....	606
15.3 Conclusion .....	614
References .....	618

**16 Optronics**

<i>J.-L. Pautrat, J.-M. Gérard, É. Bustarret, D. Cassagne, E. Hadji, C. Seassal</i> .....	619
16.1 Surface Plasmons and Nanoscale Optics.....	619
16.1.1 Introduction.....	619
16.1.2 What Is a Plasmon?.....	620
16.1.3 Dispersion Relations, Coupling with Light, and Applications.....	622
16.1.4 Optical Transmission Through Subwavelength Apertures ..	627
16.1.5 Metal Nanoparticles.....	629
16.1.6 How Far Can Plasmons Take Us?.....	633
16.2 Semiconductor Quantum Dots.....	634
16.2.1 Semiconductor Lasers: From Quantum Wells to Quantum Dots.....	634
16.2.2 Single Quantum Dots.....	640
16.3 Photonic Crystals and Microcavities.....	646
16.3.1 Introduction.....	646
16.3.2 Periodic Structures.....	646
16.3.3 Structures Without Defects. Exploiting the Allowed Bands in Photonic Crystals.....	653
16.3.4 Structures with Defects.....	656
16.3.5 Conclusion and Prospects.....	661
References.....	662

**17 Nanophotonics for Biology**

<i>J. Zyss, S. Brasselet</i> .....	665
17.1 Emission and Absorption of Light by Molecular Systems.....	676
17.2 Molecules, Supramolecular Assemblies, and Nanoparticles.....	686
17.2.1 Coupling Between Nanoparticles and Biomolecules.....	686
17.2.2 Luminescent Nanostructures Based on Semiconductors and Metals.....	695
17.2.3 Molecular Engineering for Biophotonics.....	698
17.3 Nanophotonic Instrumentation for Biology.....	707
17.3.1 Optical Detection of Single Molecules by Fluorescence.....	707
17.3.2 Multiphoton and Nonlinear Microscopy.....	731
17.3.3 Mechanical Properties of Single Biomolecules.....	737
17.4 Conclusion.....	743
References.....	746

**18 Numerical Simulation**

<i>X. Blase, C. Delerue</i> .....	749
18.1 Structural Properties.....	750
18.1.1 Interatomic Potentials and Forces.....	750



XXVIII Contents

18.1.2	Potential Energy Surface .....	752
18.1.3	Classical Molecular Dynamics .....	753
18.1.4	Monte Carlo Methods .....	755
18.2	Electron Properties .....	757
18.2.1	Basic Results from Quantum Mechanics .....	757
18.2.2	Semi-Empirical Approaches to Electron Structure .....	759
18.2.3	Ab Initio Methods .....	766
18.2.4	Ab Initio Calculation of Interatomic Forces .....	771
18.2.5	Using Electron Wave Functions and Eigenvalues .....	773
18.3	Conclusion .....	773
	References .....	774

**19 Computer Architectures for Nanotechnology:  
Towards Nanocomputing**

	<i>C. Gamrat</i> .....	777
19.1	Introduction .....	777
19.2	Computer Architecture and Basic Functions .....	779
19.2.1	Typical Architecture of a Computer .....	779
19.2.2	Memory .....	780
19.2.3	Interconnects .....	782
19.2.4	Operators .....	782
19.2.5	Technological Considerations .....	783
19.2.6	Nanomemories, Nano-operators, Nanoconnections .....	785
19.3	Some Ideas for a New Architecture .....	786
19.3.1	Calculating with Memory Alone .....	786
19.3.2	Reconfigurable Computer Architectures .....	788
19.3.3	Cellular Automata .....	789
19.3.4	Neural Networks .....	791
19.4	Computer Environment .....	794
19.4.1	Information Coding .....	794
19.4.2	Defect Tolerance .....	794
19.5	Prospects .....	798
	References .....	799
	<b>Index</b> .....	801

---

## List of Contributors

**Agnès Barthélémy**  
UMR CNRS/Thalès  
Domaine de Corbeville  
91404 Orsay, France  
agnes.barthelemy@  
thalesgroup.com

**Xavier Blase**  
Laboratoire de Physique  
de la Matière Condensée  
et Nanostructures  
Université Claude Bernard  
Lyon 1 and CNRS  
43 bd du 11 novembre 1918  
69662 Villeurbanne, France  
xblase@lpmcn.univ-lyon1.fr

**Claude Boccara**  
Ecole supérieure de physique  
et de chimie industrielles  
10 rue Vauquelin  
75231 Paris Cedex 05  
boccara@optique.espci.fr

**Jean-Philippe Bourgoïn**  
Molecular Electronics Laboratory  
Service de Physique  
de l'Etat Condensé  
CEA Saclay  
91191 Gif-sur-Yvette Cedex  
France  
jbourgoïn@cea.fr

**Arnaud Bournel**  
Institut d'électronique fondamentale  
Batiment 220  
Université de Paris Sud 11  
91405 Orsay Cedex, France  
bournel@ief.u-psud.fr

**Sophie Brasselet**  
LPQM (UMR 8537)  
and D'Alembert Institute (IFR 121)  
Ecole Normale Supérieure de Cachan  
61, Avenue du Président Wilson  
94235 Cachan, France  
sophie.brasselet@lpqm.  
ens-cachan.fr

**Pierre-François Brevet**  
Laboratoire de Spectrométrie  
Ionique et Moléculaire  
Université Claude Bernard  
Lyon 1 and CNRS  
Domaine Scientifique de la Doua  
69662 Villeurbanne Cedex, France  
pfbrevet@lasim.univ-lyon1.fr

**Etienne Bustarret**  
Laboratoire d'Etudes  
des Propriétés Electroniques  
des Solides – CNRS  
25, Avenue des Martyrs BP 166  
38042 Grenoble Cedex 9, France  
etienne.bustarret@grenoble.  
cnrs.fr

**Rémi Carminati**

Laboratoire EM2C  
CNRS, Ecole Centrale Paris  
92295 Chatenay-Malabry  
Cedex France  
remi.carminati@em2c.ecp.fr

**David Cassagne**

Groupe d'études  
des semiconducteurs  
Université Montpellier 2  
Place Eugène Bataillon  
34095 Montpellier Cedex 05  
France  
cassagne@ges.univ-montp2.fr

**Claude Chappert**

Institut d'électronique fondamentale  
Université de Paris Sud  
91405 Orsay Cedex, France  
claude.chappert@ief.u-psud.fr

**Yong Chen**

Ecole Normal Supérieure  
24 rue Lhomond  
75231 Paris, France  
yong.chen@ens.fr

Laboratoire de Photonique  
et de Nanostructures, CNRS  
Route de Nozay  
91460 Marcoussis, France  
yong.chen@lpn.cnrs.fr

**Christophe Delerue**

Institut supérieure d'électronique  
et du numérique  
41 boulevard Vauban  
59046 Lille Cedex, France  
christophe.delerue@isen.fr

**Philippe Dollfus**

Institut d'électronique fondamentale  
Université de Paris Sud 11  
91405 Orsay Cedex, France  
philippe.dollfus@ief.u-psud.fr

**Claire Dupas**

Director  
Ecole Normale Supérieure de Cachan  
61, Avenue du Président Wilson  
94235 Cachan, France  
claire.dupas@dir.ens-cachan.fr

**Joël Eymery**

CEA-CNRS Nanophysique  
et Semiconducteurs  
Building C5, Room 680  
17 Rue des Martyrs  
38054 Grenoble Cedex 9 ,France  
jeymery@cea.fr

**Arianna Filoramo**

Molecular Electronics Laboratory  
Service de Physique  
de l'Etat Condensé  
CEA Saclay  
91191 Gif-sur-Yvette Cedex  
France  
filoramo@drecam.saclay.cea.fr

**David Fraboulet**

CEA-LETI  
ST/Philips/Motorola Alliance  
850, rue Jean Monnet  
38926 Crolles Cedex, France  
fraboulet@cea.fr

**Christian Frétigny**

Ecole supérieure de physique  
et de chimie industrielles  
10 rue Vauquelin  
75231 Paris Cedex 05  
christian.fretigny@espci.fr

**Sylvie Galdin-Retailleau**

Institut d'électronique fondamentale  
Université de Paris Sud 11  
91405 Orsay Cedex, France  
Sylvie.galdin@ief.u-psud.fr

**Jean-Louis Gallani**  
IPCMS-GMO  
23 rue du Loess BP 43  
67034 Strasbourg Cedex 2  
France  
gallani@ipcms.u-strasbg.fr

**Christian Gamrat**  
CEA-LIST/DTSI/SARC/LCEI/  
Computer architectures  
CEA Saclay  
91191 Gif-sur-Yvette Cedex  
France  
christian.gamrat@cea.fr

**Jean-Michel Gérard**  
CEA Grenoble  
17, rue des Martyrs  
38054 Grenoble Cedex, France  
gerard@drfmc.ceng.cea.fr

**Marcello Fabian Goffman**  
Molecular Electronics Laboratory  
Service de Physique  
de l'Etat Condensé  
CEA Saclay  
91191 Gif-sur-Yvette Cedex  
France  
goffman@cea.fr

**Emmanuel Hadji**  
SiNaPS/CEA Grenoble  
17, rue des Martyrs  
38054 Grenoble Cedex, France  
ehadji@cea.fr

**Margrit Hanbücken**  
CRMC-N CNRS  
Campus Luminy, Case 913  
13288 Marseille  
margrit.hanbucken@crmcn.  
univ-mrs.fr

**Philippe Houdy**  
Laboratoire d'étude  
des milieux nanométriques  
Université d'Evry  
Bld F. Mitterrand  
91025 Evry Cedex, France  
philippe.houdy@univ-evry.fr

**Jean-Claude Labrune**  
Surfaces, Interfaces  
and Nanostructures  
FEMTO-ST/CREST/UMR 6174  
CNRS  
4 place Tharradin BP 71427  
25211 Montbéliard Cedex, France  
Jean-Claude.Labrune@pu-pm.  
univ-fcomte.fr

**Marcel Lahmani**  
Laboratoire d'étude  
des milieux nanométriques  
Université d'Evry  
Bld F. Mitterrand  
91025 EVRY Cedex, France  
lahmani@bp.univ-evry.fr

**Jean Lermé**  
Laboratoire de spectrométrie  
ionique et moléculaire  
Université Claude Bernard  
Lyon 1 and CNRS  
Domaine Scientifique de la Doua  
69662 Villeurbanne Cedex, France  
lerme@lasim.univ-lyon1.fr

**Annick Loiseau**  
Laboratoire d'Etude  
des Microstructures (LEM)  
UMR 104 Onera-CNRS  
ONERA, BP 72  
29 avenue de la Division Leclerc  
92322 Chatillon Cedex, France  
Annick.Loiseau@onera.fr

**Dominique Mailly**

Laboratoire de photonique  
et des nanostructures  
route de Nozay  
91460 Marcoussis, France  
dominique.mailly@lpn.cnrs.fr

**Patrice Mélinon**

Laboratoire de Physique  
de la Matière Condensée  
et Nanostructures  
Université Claude Bernard  
Lyon 1 and CNRS  
Domaine Scientifique de la Doua  
69662 Villeurbanne Cedex, France  
patrice.melinon@lpmcn.  
univ-lyon1.fr

**Jean-François Nierengarten**

Groupe de Chimie des Fullèrenes  
et des Systèmes Conjugués  
Laboratoire de Chimie  
de Coordination du CNRS  
205 route de Narbonne  
31077 Toulouse Cedex 4, France  
jfnierengarten@lcc-toulouse.fr

**Frank Palmino**

Surfaces, Interfaces  
and Nanostructures  
FEMTO-ST/CREST/UMR 6174  
CNRS  
4 place Tharradin BP 71427  
25211 Montbéliard Cedex, France  
Frank.Palmino@pu-pm.  
univ-fcomte.fr

**Jean-Noël Patillon**

MOTOROLA Labs  
Espace Technologique Saint-Aubin  
91193 Gif-sur-Yvette Cedex  
France  
Jean-Noel.Patillon@Motorola.  
com

**Jean-Louis Pautrat**

CEA-Grenoble and Minatec  
17, rue des Martyrs  
38054 Grenoble Cedex, France  
pautratjl@chartreuse.cea.fr

**Anne Pépin**

Laboratoire de Photonique  
et de Nanostructures, CNRS  
Route de Nozay  
91460 Marcoussis, France  
pepin@lpn.cnrs.fr

**Alain Perez**

Laboratoire de Physique  
de la Matière Condensée  
et Nanostructures  
Université Claude Bernard  
Lyon 1 and CNRS  
Domaine Scientifique de la Doua  
69662 Villeurbanne Cedex, France  
alain.perez@lpmcn.  
univ-lyon1.fr

**Sylvie Rousset**

Group de Physique des Solides  
Universities of Paris 6 and 7  
CNRS, 2 Place Jussieu  
75251 Paris Cedex 5, France  
rousset@gps.jussieu.fr

**Yves Samson**

DSM/CEA Grenoble  
17, avenue des Martyrs  
38054 Grenoble Cedex 09  
France  
ysamson@cea.fr

**Christian Seassal**

LEOM-UMR CNRS 5512  
Ecole centrale de Lyon  
Batiment 7  
36 Avenue Guy de Collongue  
BP 163, 69131 Ecully Cedex  
France  
Christian.Seassal@ec-lyon.fr

**Nathalie Solladié**  
Groupe de Synthèse  
de Systèmes Porphyriniques  
Laboratoire de Chimie  
de Coordination du CNRS  
205 route de Narbonne  
31077 Toulouse Cedex 4  
France  
solladie@lcc-toulouse.fr

**Didier Stiévenard**  
IEMN Lille  
41 boulevard Vauban  
59046 Lille Cedex, France  
didier.stievenard@isen.fr

**Christophe Vieu**  
LAAS/CNRS  
7 avenue du Colonel Roche  
31077 Toulouse Cedex 4  
christophe.vieu@laas.fr

**Dominique Vuillaume**  
IEMN/CNRS Lille  
BP 69, Avenue Poincaré  
59652 Villeneuve d'Ascq  
Cedex, France  
vuillaume@iemn.univ-lille1.fr

**Joseph Zyss**  
LPQM (UMR 8537)  
and D'Alembert Institute (IFR 121)  
Ecole Normale Supérieure de Cachan  
61, Avenue du Président Wilson  
94235 Cachan, France  
joseph.zyss@lpqm.ens-cachan.fr

## **Part I**

---

### **Tools for Nanoscience**

# Lithography and Etching Processes

D. Maily and C. Vieu

## 1.1 Definitions and General Considerations

Lithography is the process of printing patterns onto a thin film called a resist, using a localised interaction between this layer and an engraving micro-tool or particle beam.

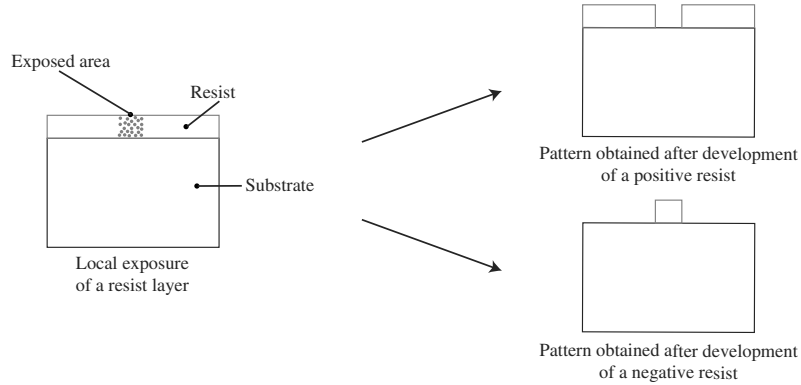
The various techniques of lithography can be classified according to the micro-tool or the type of radiation used (detailed in Sect. 1.5). Hence, to print the pattern, photolithography uses photons, electron lithography uses electrons, and ion lithography uses ions. On the other hand, lithography by impression uses the mechanical interaction between a hard mould and a layer of soft resist, and near-field lithography uses various types of interaction (electrical, mechanical, thermal, optical) between a fine tip and the surface of the resist.

The lithography itself does not therefore structure the active material which will constitute the core of the nanodevice. It simply sketches the outline of the future device in a sacrificial layer, the resist, and this is then used in a transfer stage to shape the active layer according to the dimensions of the pattern imposed in the lithography stage.

## 1.2 Photoresists

The photoresist is a thin layer deposited on the surface of the active material destined to receive the radiation or the interaction used during the lithographic process. The word ‘photoresist’, or ‘resist’ for short, is used for historical reasons. Optical lithography, or photolithography, which was the precursor of all modern microlithographic techniques, used a polymerised organic material or resin for this purpose.





**Fig. 1.1.** Local exposure of resist and development

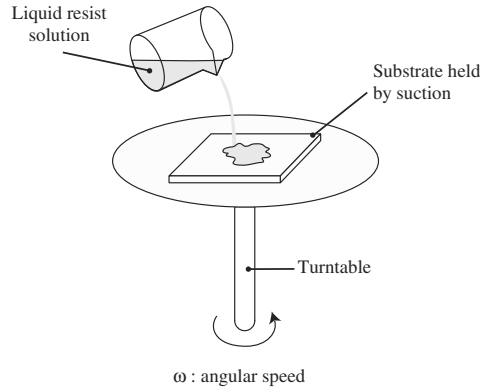
## Lithographic Materials

With the development of a great many micro- and nanolithographic methods, other types of material are now used to print patterns. In electron lithography, for example, inorganic films such as  $\text{AlF}_3$ ,  $\text{SiO}_2$  and  $\text{MgO}$  can be exposed. In STM (scanning tunneling microscopy) or electron beam lithography, patterns can be printed on self-assembled monolayers (SAM). We shall also see in Chap. 3 how transferable patterns can be printed on the passivation layer of a hydrogenated silicon surface using an STM tip. In the latter case, the resist layer is in fact the uppermost atomic layer of the surface containing hydrogen atoms which attach to the dangling bonds of a monocrystalline silicon surface. It is clear from this that the word ‘photoresist’ is intended to mean a sacrificial layer on which a pattern can be printed. Figure 1.1 shows a process using a standard resist, which serves to illustrate the general approach. There are many variations on this theme, depending on the chemical or structural characteristics of the resist layer and the nature of the specific interaction used for lithography.

### 1.2.1 Example of Processing with a Polymer Resist

A polymer resist is a typical photoresist for photolithography or electron lithography. The resist is an intelligent polymer comprising two parts: a matrix, insensitive to the writing radiation, which fulfills the mechanical requirements of the resist, and an active component, sensitive to the radiation, which either accelerates or slows down the rate at which the resist dissolves in a solvent. There are thus two types of resist: positive resists for which exposure increases the solubility and negative resists for which exposure reduces the solubility.

- The polymer constituting the resist is dissolved in a solvent to obtain a liquid.
- The substrate is coated with resist on a turntable (spin-coating). The thickness of the resist coating can be very accurately controlled, to within



**Fig. 1.2.** Spreading the resist layer using a turntable

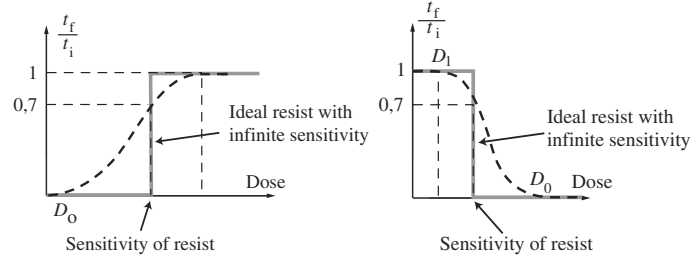
a few nanometers, by adjusting the solubility of the polymer in the solvent, the intrinsic viscosity of the polymer macromolecules, and the angular speed of the substrate on the turntable.

- Before exposure, the substrate is raised to a moderate temperature (around  $100^{\circ}\text{C}$ ) to evaporate any excess solvent molecules incorporated within the resist layer (soft bake). This makes the thickness of the coating more uniform.
- The resist is exposed using the lithography tool. The resist layer is modified locally. In general, these modifications are of a chemical nature and no topographical features are visible on the layer. A latent image is formed at the end of the lithography process.
- The result is then developed by immersing the substrate in the appropriate solvent, which dissolves the resist selectively according to the degree of exposure. In the case of a positive resist, development leads to the formation of a hole in the exposed regions (weak solvent for the initial polymer), whereas in the case of a negative resist, development dissolves the resist rather in the unexposed regions (strong solvent for the initial polymer).
- Finally, there is a post-exposure bake, in which the substrate is raised to a higher temperature (around  $120^{\circ}\text{C}$ ) in order to evaporate excess solvent molecules from the development phase, and in some cases to harden the polymer in an irreversible manner by favouring cross-link reactions between the macromolecular chains. Any resist remaining on the surface is then more robust for the ensuing microfabrication operations.

### 1.2.2 Sensitivity and Contrast

The most important parameters characterising a resist layer are sensitivity and contrast.

The sensitivity of the resist refers to the intensity of the relevant radiation or interaction used in the lithography, often called the dose, required to



**Fig. 1.3.** *Left:* Negative resist. The resist hardens during exposure and the contrast curve increases. *Right:* Positive resist. The resist is softened during exposure and the contrast curve decreases

cause a sufficient modification of the resist to ensure that the desired pattern appears at the development stage (when such is necessary). This parameter is analogous to the sensitivity of a photographic film. Naturally, it is the sensitivity of the resist that determines the total length of exposure. In industrial lithography where mass production is imperative, highly sensitive resists are preferred.

The sensitivity of a resist is expressed in units characterising the type of interaction used for lithography. When charged particle beams are used to irradiate the resist (electron or ion lithography), typical units are coulomb/cm<sup>2</sup>. When the resist is irradiated by a photon beam, typical units are J/cm<sup>2</sup>.

The contrast of the resist characterises the variation of the solubility rate in its developer as a function of the exposure time (dose). The higher the contrast, the better the resist will be able to reveal small variations in the received dose. This is a crucial feature of the resist. Indeed, as we shall see later, whatever type of lithography is used, the spatial localisation of the exposure on the resist never cuts off abruptly. Owing to various physical effects that depend on the type of radiation or interaction used (diffraction for photons, collisions for electrons, etc.), the actually exposed region of the resist extends slightly beyond the intended patterns to include a transition zone that varies in width. These transition zones determine the spatial resolution of the lithography process. It is intuitively clear that, the higher the contrast of the resist, the less these edge effects will contribute to spreading of the patterns. It is therefore a priority to find high-contrast resists.

Note, however, that it is an abuse of language to speak of the contrast of the resist, because this contrast is only defined for the complete lithography process, which involves not only the resist, but also the type of radiation or interaction used, the developing solution and the developing temperature. Contrast curves are generally obtained experimentally for this set of parameters.

### Contrast Curves of a Resist

The final thickness of the resist film is measured experimentally after development and compared with the initial thickness of the resist film after deposition, to give the parameter  $t_f/t_i$ , for various values of the radiation dose.

For a positive resist,  $D_0$  denotes the threshold dose beyond which the final resist thickness becomes unmeasurable, whilst  $D_1$  denotes the threshold dose below which the final resist thickness does not differ significantly from the initial thickness before exposure. The contrast  $\gamma$  of the resist is a measure of the steepness of the contrast curve between  $D_0$  and  $D_1$ :

$$\gamma = (\log D_0 - \log D_1)^{-1}.$$

#### 1.2.3 Example of a Positive Resist

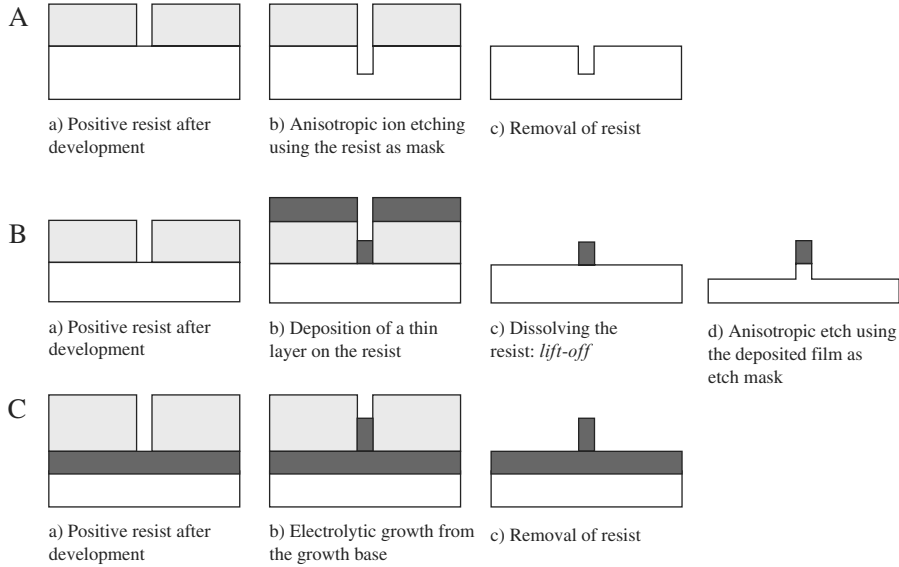
For concreteness, let us examine a typical example of a positive resist commonly used in electron lithography, namely polymethylmethacrylate (PMMA), better known by the generic name of plexiglass. This polymer is generally used with a very high molecular weight of something like a million. Once the resist film has been spread, it forms a dense network of enormous macromolecules with a high level of entanglement. The effect of an electron beam, or more generally, ionising radiation, is to trigger a rather complex set of chemical reactions which break carbon-carbon bonds in the polymer backbone. Hence, the main effect of irradiating the resist is a local reduction in the molecular weight of the polymer. In the region exposed during the lithographic process, the network of polymer macromolecules is loosened and the chains become less entangled. Now it is well known that the effect of a solvent on a polymer depends sensitively on the molecular weight of that polymer. Indeed, since the solvent molecules must penetrate within the macromolecular network, this penetration will clearly be enhanced when the molecular weight is reduced. Choosing a solvent that is well-suited to PMMA, it is thus possible to dissolve the exposed regions in a selective manner, without disturbing those regions that have been protected from irradiation. A ‘hole’ is then formed at the site of the irradiated patterns.

The initial and final molecular weights  $M_i$  and  $M_f$  of the polymer are simply related:

$$M_f = \frac{M_i}{1 + g\varepsilon M_i/\rho A_0},$$

where  $g$  is the number of broken bonds per unit energy absorbed by the resist during exposure,  $\varepsilon$  is the energy deposited per unit volume during exposure,  $\rho$  is the density of the resist, and  $A_0$  is Avagadro’s number.

According to this expression, the parameter  $g$  determines the sensitivity of the resist. This simple expression also shows that, if one is able to calculate the spatial distribution of energy deposited during exposure of a pattern, then one can account for the evolution of the molecular weight of the polymer film



**Fig. 1.4.** Three examples of transfer strategies starting from a pattern obtained by lithography on a positive resist

at any point. For example, in electron lithography, the possible trajectories of incident, back-scattered and secondary electrons can be simulated. It is then possible to calculate the energy contributed by these electrons in inelastic interactions with the resist atoms. This deposited energy and its spatial distribution determine the size and shape of the pattern obtained in the resist after development. If this equation is combined with a simple law characterising the selective solubility of the polymer in the developer solvent, it is in principle possible to predict the size and shape of the patterns. An empirical law of the type

$$V = V_0 + \beta/M_f^\alpha$$

generally gives good results. In this equation  $V$  is the solubility rate of the resist film with final molecular weight  $M_f$  in the developer solvent,  $V_0$  is this solubility rate for an infinite molecular weight (approximately that of the unexposed resist), and  $\beta$  and  $\alpha$  are parameters to be determined by simple calibration experiments.

### 1.2.4 Transfer Stage

We have just seen how a resist layer can be used to produce patterns when irradiated by a particle beam. This lithographic process is of course a key stage of nanofabrication, but it is far from sufficient to satisfy all needs. Indeed, the resist itself is often not the material in which nanostructures are to be created,

but merely constitutes a sensitive sacrificial layer. The patterns printed in this resist layer must then be transferred to the relevant material. As far as possible, this transfer stage, just as crucial as the lithographic stage itself, must preserve the size and shape of patterns drawn in the resist. Figure 1.4 shows schematically several transfer techniques used with a positive resist. In the following, without seeking to provide an exhaustive discussion, the aim will be to explain the main strategies used to convert a resist pattern into a functional nanostructure.

### 1.3 Subtractive Pattern Transfer

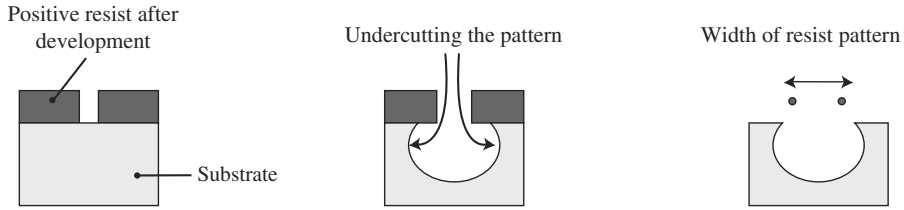
In the so-called subtractive transfer technique, the idea is to use patterns printed in the resist layer to etch the sample surface solely in those regions stripped of resist after development.

#### 1.3.1 Wet Etching

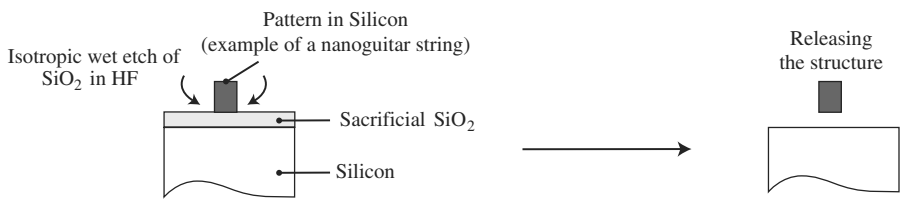
##### Basic Principle

The sample surface is etched chemically by immersing it in a solution containing reactants specific to the substrate but inert with regard to the material used as a mask (e.g., the resist layer after exposure). The advantages with this approach are the ease with which it can be implemented, the wide range of etching solutions for every type of material, and above all the speed of the process, which, depending on the concentration of reactive elements in the solution, can be very high indeed (several microns per minute). The main disadvantage which disallows use of this method in the vast majority of cases when nanometric patterns are to be etched is that it acts isotropically. As shown in Fig. 1.5, the etch front moves isotropically, i.e., the surface is etched in all directions within the pattern and this leads to a considerable broadening of the pattern after etching. The lateral dimensions of the structures are no longer precisely controlled. However, it should be noted that the isotropic character of the etch is sometimes used deliberately to free structures from their substrate. Nanostructures are in fact undercut by etching a sacrificial  $\text{SiO}_2$  layer isotropically (see Fig. 1.6).

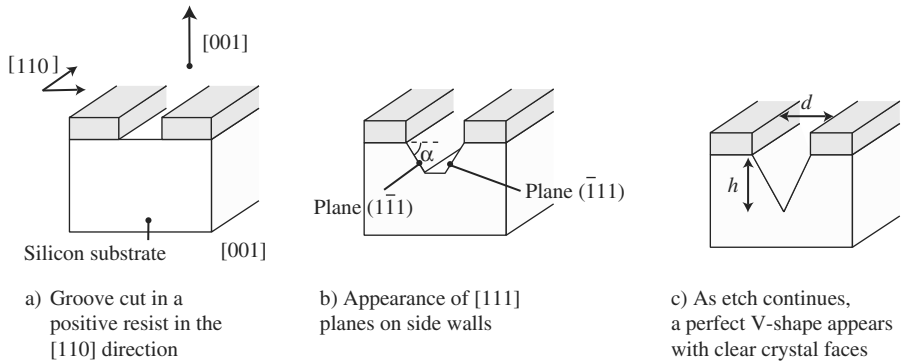
Wet-etching of monocrystalline materials can sometimes be distinctly anisotropic. This is because the etch rate can be very different for different crystallographic planes of the material. The best known and most widely used example is monocrystalline silicon. For this material there exist certain alkaline etchant solutions that have almost no effect on the dense planes of type  $\{111\}$  in the diamond structure of the silicon. Hence, if the patterns are suitably oriented with respect to the crystal axes, quite remarkable profiles can be developed, revealing certain atomic planes of the structure and sometimes limiting undercut effects like those shown in Fig. 1.7.



**Fig. 1.5.** Isotropic effect of a wet etch, e.g., when silicon is etched with a mixture of nitric acid, hydrofluoric acid and water



**Fig. 1.6.** Undercutting a nanostructure by wet-etching a sacrificial SiO<sub>2</sub> layer with a mixture of hydrofluoric acid and water. The endpoints of the freed structure shown here in cross-section rest on the substrate at anchoring points that have not been represented



**Fig. 1.7.** Anisotropic wet etch of monocrystalline silicon [001] in an etchant of type KOH + water. The etch rate of this solution is negligible along the planes {111}. With this configuration, a V-shaped profile is etched out when the pattern edges are aligned with a surface of type [110], and undercutting effects remain minimal. The angle  $\alpha$  is about  $54^\circ$  ( $\cos \alpha = 1/\sqrt{3}$ ). The depth  $h$  of the V-shape and the width  $d$  of the top of the V-shape are thus related by  $2h = d\sqrt{2}$

Many ingenious systems have been devised to create quite novel structures using crystallographic effects. In the field of nanotechnology, this process has been widely used to pattern semiconductor substrates (Si, GaAs, InP, etc.) before epitaxial deposition of thin nanometric films. However, the restriction to monocrystalline materials and the constraint imposed by the specific

crystalline structure of the material mean that this elegant technique cannot be generalised to a wide range of applications.

### Advantages and Disadvantages of the Technique

To sum up then, wet etching is a very simple process with good etch rates and a high level of selectivity between different types of material. The latter feature, related to the chemical nature of the etch, is fundamental whenever one needs to curtail the etch instantly at a given depth. In that case, it suffices to insert a stopping layer, inert with regard to the etchant, into the substrate at the required depth. Chemical selectivity is also an advantage when it comes to choosing a mask that can resist the effects of the etchant. There is a great deal of literature (see for example [1]) and even encyclopedic resources listing chemical etchants for a wide range of materials, with information such as their isotropic or anisotropic characteristics, etch rate, and suitable choices of inert materials to use as etch masks.

It should nevertheless be noted that the use of wet etching for nanometric patterns remains extremely limited due to undercut effects which make it difficult to control the lateral dimensions of target structures and in particular to obtain a truly vertical profile in etched patterns.

### 1.3.2 Dry Etching

#### Basic Principle

In this approach, the sample is etched by bombarding the surface with high-energy ions (several tens of eV to several keV) in a vacuum environment. It has long been known that elastic collisions between incident ions and surface atoms can cause a great many of those atoms to be removed from the material. This ion erosion phenomenon is known as sputtering. The efficiency of ion removal is characterised by the sputtering yield  $S$ , which stands for the number of ejected atoms per ion incident at the surface. A relatively simple expression for this parameter due to Sigmund is

$$S = \frac{3}{4} \frac{E_d}{N\pi^2CU}, \quad (1.1)$$

where  $C = 1.81 \text{ nm}^2$  is a constant,  $N$  is the atomic density of the material (in atoms/cm<sup>3</sup>),  $U$  is the binding energy of surface atoms (e.g., 6 eV for silicon), and  $E_d$  is the energy deposited in an elastic collision when an ion is incident on the material surface. The latter quantity characterises the collision efficiency of the incident ion with regard to the target material. It is quite straightforward to calculate from the stopping powers predicted by the collision cross-section for the ion and target atoms. The unit used for this parameter is generally eV/nm because of its relationship with a stopping power (energy given up per unit matter crossed). Typical values are of the order of a few tens of eV/nm.



The deposited energy  $E_d$  thus depends on the energy of the incident ions, the type of ions and the atoms in the material (atomic mass and number), and the angle between the incident ions and the surface (the angle of incidence). Concerning the latter, grazing incidence tends to favour sputtering because the ions penetrate less deeply and thus deposit more energy in the surface layers.

Sigmund's formula is therefore easy to interpret: the more tightly the atoms of the material are bound to the surface (large  $U$ ), the lower the sputtering yield will be; the more efficiently each ion transfers energy to surface atoms (large  $E_d$ ), the higher the sputtering yield will be. Standard values of the sputtering yield for typical ions such as argon accelerated to a few keV are of the order of 5–10. It should be no surprise to find that these values are greater than unity. Indeed, an incident ion creates a great many collisions within the material, thereby displacing a large number of atoms which can in turn generate further (secondary) collisions. This proliferation of generated collisions, commonly known as a cascade, explains why a single ion is able on average to strip a number of atoms from the sample surface.

### Advantages and Disadvantages of the Technique

The simplest and archetypal dry etching method is ion beam etching (IBE). In this process, an ion beam is directed with normal incidence at the sample surface. (This beam can be electrostatically neutralised to avoid charge effects in the target material.) Ions are typically noble gas ions such as argon, which exhibit no chemical activity with respect to the target atoms. Etching is therefore purely collisional and one refers to this as physical rather than chemical etching, in contrast to what happens in wet etches.

The primary quality of this type of etch is that it produces almost vertical sides on etched features, due to the normal incidence of the ions on the sample surface and their large kinetic energy in the perpendicular direction.<sup>1</sup> This means that the lateral dimensions of patterns can be preserved during the etch. This is the main reason why the vast majority of etched nanostructures are obtained by dry etching. (The gates in today's commercially produced CMOS transistors are manufactured by dry-etching a film of polycrystalline silicon.)

On the other hand, purely physical IBE-type dry etching is not without drawbacks of its own. For one thing, it is slow. Ion sources can be made to produce current densities of the order of 1 mA/cm<sup>2</sup>, and when this is multiplied by a typical sputtering yield, etch rates of the order of a few tens of nanometers per minute are obtained. For another thing, it is non-selective, since all

<sup>1</sup> The side walls of etched features can never be perfectly vertical owing to gradual erosion of the etch mask and redeposition of sputtered material on the side walls. However, these unwanted effects can be significantly reduced by tilting the sample through several degrees with respect to the incident ion direction and rotating it about its normal to avoid shadow effects from the mask.

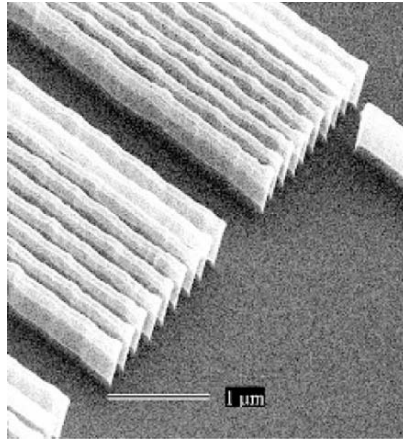
materials are eroded by ion bombardment. Even though the sputtering yield may vary from one material to another, such variations remain small. This second fact explains why the etch mask is itself sputtered during the etch. For example, when the mask is formed from the resist remaining after lithography and development, the effect is then dramatic, since the resist has a rather low atomic density  $N$  and so will be sputtered even more efficiently than the substrate [see the expression (1.1) for the sputtering yield  $S$ ].

To sum up, dry ion etching with inert ions has all the advantages that wet etching does not, being highly anisotropic, whence the side walls of etch features are almost vertical, but it also has all the drawbacks that wet etching avoids, being slow and non-selective. On the basis of this conclusion, the idea was born to combine in a single etch system a chemical component, using species that react strongly with the surface, and a physical component, using ion bombardment, so as to unite speed, selectivity and anisotropy.

### 1.3.3 Reactive Ion Etching

This is the general idea behind reactive ion etching (RIE), which is today by far the most widely used etch process to transfer nanometric patterns. To this end, a plasma radio frequency (13.56 MHz) is created inside a chamber which has been evacuated and then filled with a gas mixture containing molecules that will generate radicals chosen to react with the sample surface. The latter is placed at the cathode of the system, which is generally coupled capacitively to the RF generator. This setup is designed so that it will spontaneously generate a negative potential at the sample surface when the plasma is initiated, due to the much higher mobility of free electrons in the plasma compared with the ions. It is this ‘self-polarising’ potential that accelerates positive ions in the plasma towards the sample and hence causes ion sputtering of the sample surface.

The idea behind reactive ion etching is very elegant. When the RF plasma is initiated, the gas precursors dissociate into a great many chemical species. Amongst these are certain electrically neutral radicals which are chemically highly reactive with respect to the sample surface. These reactive radicals form highly volatile compounds at the sample surface. At the same time, ions and electrons are produced in great numbers and the growing negative potential at the surface sets the ion bombardment in motion. In this way, chemical and physical etching are brought about in synergism. The reactive radicals considerably reduce the binding energy  $U$  of surface atoms and ion bombardment thus strips the surface with a high sputtering yield [see Sigmund’s formula (1.1)]. The ingenuity of the operator then goes into judicious adjustment of the plasma parameters (type of gas injected, gas pressure, RF power) so as to achieve a highly anisotropic etch with side walls as near to vertical as possible, whilst activating a surface chemistry that procures the desired selectivity between materials and high etch rates.



**Fig. 1.8.** Silicon walls etched by anisotropic RIE. Note the verticality of the sides. The walls have width 30 nm and height 600 nm, giving an aspect ratio of 1/20. The SEM (scanning electron microscope) image was taken at an angle to show that the very thin walls are transparent to the electron beam of the microscope. The brighter part at the top of the pattern is the metallic etch mask, in this case a 10-nm chromium film. Photo F. Carcenac (LAAS/CNRS)

Figure 1.8 shows a particularly representative example of a silicon nanostructure produced by highly anisotropic RIE. Dry etching by plasmas is a field of intense study today. Research aims to develop ways of analysing the plasma, spectroscopic analysis of the chemical radicals involved, and surface characterisation, in order to obtain a better understanding of all the mechanisms brought into play and optimise operating conditions. The goal here is to combine the (isotropic) chemical etch and the (anisotropic) physical etch with as high a level of control and reproducibility as possible.

There exist many different plasma etch processes able to produce deep vertical features in a wide range of materials, as in the example shown in Fig. 1.8. The most advanced processes use passivation layers formed on the sample surface by plasma-assisted polymerisation of carbon-bearing species. These films inhibit the etch wherever they cannot be stripped by the ion bombardment. The side walls of patterns, protected by such passivation layers and not directly exposed to the ion bombardment, which arrives perpendicularly to the surface, remain intact during the etch, thereby leading to vertical etch profiles.

As in the case of wet etching, there is a vast literature and whole libraries about plasma etch processes where those wishing to create nanometric structures in some specific material can obtain advice on the choice of gas mixtures and proportions, plasma parameters, suitable materials for the etch mask, and so on.

## 1.4 Additive Pattern Transfer

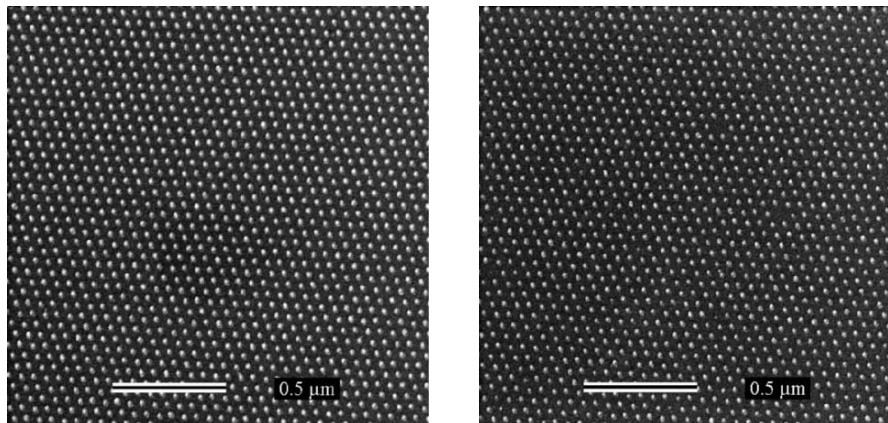
Figure 1.4 (B and C) illustrates two additive transfer techniques, in which the aim is to exploit openings made in the resist film during lithography in such a way as to deposit a new material on the sample surface. This material can be deposited by what are usually called physical methods, such as vacuum vapour deposition or sputtering. The technique used to locate this deposition precisely in the openings made in the resist is called lift-off (see Fig. 1.4B). One can also use resist patterns as a mould for growth via an electrochemical reaction in a liquid medium, which deposits electrolyte ions on the surface that remains unprotected by the resist. This is called electrolytic growth transfer (see Fig. 1.4C).

### 1.4.1 Lift-Off

A thin film of the material to be transferred is deposited on the surface of the resist after the lithographic process, ensuring that the deposit formed above the resist and inside the patterns is discontinuous across the patterns. The resist layer is then simply dissolved in a suitable solvent, whereupon the layer of material deposited on top of the resist is removed, leaving only that part of the deposit located at the openings in the resist, i.e., in contact with the sample surface. The result is that the material is only deposited within the patterns originally printed in the resist. The openings in the resist are thus transformed into a deposited pattern of the chosen material localised on the nanometric scale. The lift-off process is very simple and efficient. A fine example is shown in Fig. 1.9. An array of nanosized dots has been obtained by lift-off of a thin platinum film after electron lithography, for applications in catalysis.

Successful implementation of the lift-off process is a matter of simple common sense. To begin with, the film deposited on the resist must be clearly discontinuous across resist features. For this purpose, the lithography parameters must be adjusted to yield highly vertical, or even slightly overhanging side walls in the resist. This reduces the risk of depositing material on the side walls. Secondly, one needs a highly directional deposition method (typically vapour deposition in vacuo), once again to avoid depositing material on the side walls.

Naturally, deposition imposes very tight constraints. Indeed, lift-off can only work if resist patterns are unaffected by the deposition. In particular, if deposition is carried out at a temperature above the glass transition temperature of the resist, those patterns will be obliterated. In most lift-off processes, deposition is carried out at room temperature, which is often incompatible with the requirements of epitaxy. The deposited material is then either amorphous or polycrystalline, depending on the nature of both the deposited material and the substrate.



**Fig. 1.9.** Pt nanoparticle array fabricated by electron lithography and lift-off of a 6-nm Pt film. The Pt particles of diameter 10 nm (*left*) and < 10 nm (*right*) are arranged in a lattice of period 50 nm. Photo F. Carcenac (LAAS/CNRS)

Moreover, the thickness of the deposited layer must be less than the thickness of the resist layer, otherwise the patterns may be completely blocked up with deposited material. In practice, this leads to a rather well established empirical rule: for very small patterns (less than 20 nm), it is almost impossible to deposit a layer of material with thickness greater than the lateral dimension of pattern features. It should thus be remembered that this additive transfer technique can never be used to deposit thick layers, and this all the more so as the patterns become smaller. Lift-off is the most common transfer technique when electron lithography is used to fabricate different types of nanodevice (one-electron transistors, micro-squids, etc.), or to define a hard mask on the sample surface for a subsequent etching stage (as in the example of Fig. 1.8).

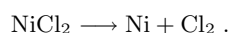
#### 1.4.2 Electrolytic Growth

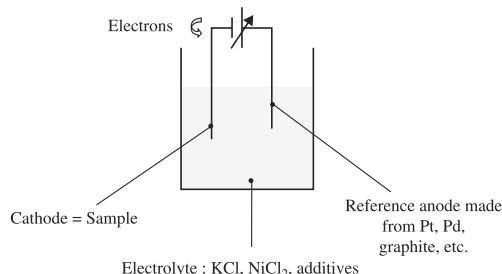
The idea of using electrolytic growth was invented mainly to obtain thicker patterns than those produced using lift-off.

##### Example of Nickel Deposition

The idea here is to carry out a redox reaction in an electrolytic cell. The cell functions as a receiver. An external generator then forces a current through the cell and thereby controls the kinetics of the redox reaction.

At the cathode,  $\text{Ni}^{2+}$  ions in the electrolyte are reduced ( $\text{Ni}^{2+} + 2e^- \rightarrow \text{Ni}$ ) and deposited on the sample surface, whilst at the anode,  $\text{Cl}^-$  ions are oxidised ( $2\text{Cl}^- \rightarrow \text{Cl}_2 + 2e^-$ ), and chlorine gas given off. The overall redox reaction is thus





**Fig. 1.10.** Electrolytic cell for nickel deposition

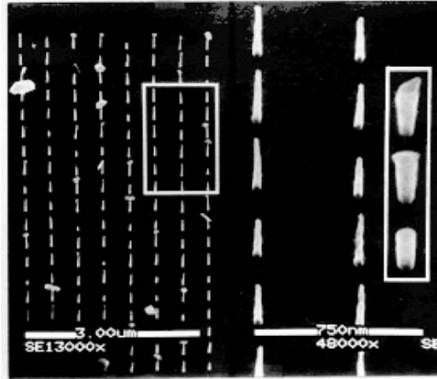
The openings in the resist are then used as a mould to localise the deposit within the patterns defined during lithography. Figure 1.4C illustrates this process schematically. In general, if the substrate is not a good conductor, a thin metal film must be deposited under the resist to serve as a base for electrolytic growth and provide a good electrical contact on the sample surface. This metal film is generally chosen to have a very strong binding to the substrate surface. Electrolysis is carried out in a quite conventional manner in an electrolytic cell. The sample is placed at the cathode of the system and a reduction reaction occurs there, wherein ions from the electrolyte are deposited in proportion to the electrical charges exchanged with the external generator. The thickness of the deposit is very simply controlled by adjusting the polarisation current and growth time and using Faraday's law. This gives the mass of metal deposited as

$$m = \frac{ItM}{Fz} ,$$

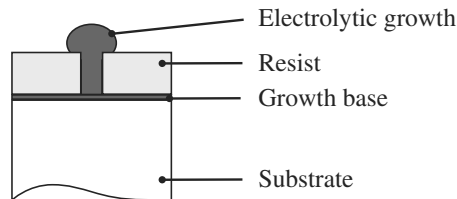
where  $I$  is the total current intensity from the generator,  $t$  is the growth time,  $M$  is the molecular mass of the deposited material,  $F$  is the Faraday charge unit (96 500 C), and  $z$  is the valence of the electrolyte ions.

This deposition technique, often used in industry, is very easy to implement. Moreover, it is highly reproducible and low in cost, and has been used in the nanotechnology field for several years now. In particular, it has made it possible to obtain metallic multilayers of very high structural quality, comparable with those obtained by epitaxy in ultrahigh vacuum. One advantage of this method is that thick layers can be deposited very quickly. However, when transfer is carried out by electrolytic growth inside patterns of nanometric dimensions, certain new effects arise:

- The electrolyte is more difficult to regenerate, so large pattern features grow more quickly than narrow ones.
- The growth rate varies from one pattern to another because it depends on the crystallographic orientation of the crystal grains in the growth base, and if the latter is polycrystalline without texturing, each pattern statistically samples all possible crystal orientations on the surface.



**Fig. 1.11.** Gold pillars obtained by electron lithography and electrolytic growth. The pillars have diameter 30 nm. In these tiny nanostructures, the growth rate from one object to another is not the same. Some pillars are only 50 nm high, whereas others exceed 300 nm. In the latter case, electrolytic growth has overflowed the resist mould and a typical mushroom shape is observed. Photo A.M. Haghiri (IEF)



**Fig. 1.12.** Typical shape of a pattern feature obtained when electrodeposition has overflowed the resist mould. This mushroom shape has been deliberately used to make gates for ultra-high speed transistors

The effects specific to the nanoscale are well illustrated by the example in Fig. 1.11. The additive pattern transfer after electron lithography is achieved here by electrodeposition in a pattern of 30-nm dots. It is quite clear that, although electrolytic growth is possible within such small features, the pillars produced in this way do not all have the same height. Some are very small, while others reach the total thickness of the resist (300 nm in this example), and still others have grown much more quickly and ended up overflowing the resist mould to produce a kind of mushroom formation at the top (see Fig. 1.12). However, it remains true that the technique is capable of depositing thick layers in nanometric features. One can thus obtain structures with a high aspect ratio (300 nm high and 30 nm across in this case), which are inaccessible using lift-off techniques. Electrolytic growth is therefore preferred to lift-off as an additive transfer method when one needs to combine small dimensions with large thicknesses. This is the case when making masks for X-ray lithography.

## 1.5 Lithography

### 1.5.1 Overview of Lithographic Methods

Any lithography method can be characterised by the type of interaction used to modify the resist layer. Hence, one speaks of optical lithography when some form of electromagnetic radiation is used to expose the sample, or electron lithography when an electron beam is the writing tool. The other important characteristic of any lithographic technique is the way patterns are written on the resist. There are two main families of techniques: parallel writing methods and sequential writing methods.

In the first category (parallel writing), the whole pattern is made simultaneously using a mask which dictates the features to be reproduced. This is analogous to the projection of a transparency by an overhead projector. The transparency carries the message to be projected and thus plays the role of the mask. It is placed on the projector and its contents are reproduced instantaneously as a single image on the screen. It is easy to understand then that parallel lithography techniques are faster, since a whole chip can be exposed in a single stage. The price to pay is that one must first make the mask, containing all the patterns that need to be reproduced. The mask serves as a template that can be reused a great many times. Strictly speaking, these lithography techniques are therefore just methods for duplicating a mask.

In sequential lithography techniques, patterns are written point by point on the resist surface. This is analogous to writing a message on a blackboard, letter by letter, with a piece of chalk. Pattern features are formed using a basic tool which exposes the resist film pixel by pixel. It is quite clear that these techniques are much slower. On the other hand, there is absolutely no need to produce a mask as template for implementing the lithography stage.

In the mass production industry, parallel duplication methods are preferred for the actual production process due to their high yield, whilst the masks required for these processes are made by the slower sequential techniques, because they tend to be more precise. The two main parameters of a lithography technique are:

- its resolution, i.e., the size of the smallest pattern feature that can be fabricated,
- its writing speed, i.e., the area that can be exposed per unit time.

We shall see in the next section that, unfortunately, these two parameters are hard to reconcile. The very fast parallel techniques, such as optical lithography, often result in poor resolution ( $> 50$  nm), limited by diffraction effects. On the other hand, sequential methods like electron beam lithography with very high resolution ( $< 10$  nm) involve very slow writing speeds.

It is this unfortunate state of affairs which means that we still do not have a lithographic technique capable of mass-producing nanometric structures (of a few nanometers). This is the main obstacle in current nanotechnology, responsible for the fact that certain nanocomponents cannot yet be commercialised



in applications for the general public. New technological tools now under investigation in the laboratory, such as AFM lithography (using atomic force microscopy), nanoimprinting, EUV lithography (using extreme ultraviolet radiation), soft lithography, and others, aim precisely to solve the problem of reconciling resolution and speed. These technologies, still in research, cannot yet be used for large scale mass production.

### 1.5.2 Proximity and Contact Photolithography

#### Basic Principle

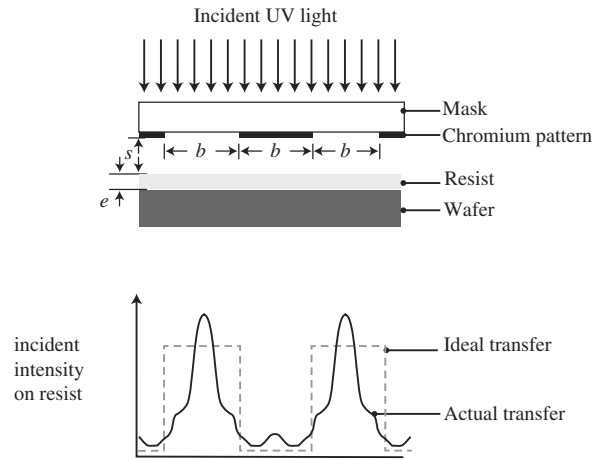
Proximity and contact lithography are the oldest methods used to reproduce a pattern by ultraviolet light. A wafer is coated with a photosensitive resist and exposed to UV light via a mask which is held against it or in close proximity. The mask has opaque and transparent parts which reproduce the relevant pattern. It is generally a quartz plate coated with chromium in those regions where opacity is required. This very simple technique represented the mainstay of microfabrication until the mid-1970s.

The resolution limit with this approach is due to light diffraction at the edges of the opaque regions. To discover the spatial distribution of the light intensity very near an edge, one cannot use the Fraunhofer diffraction theory, which is only valid in the far field. One must therefore use the Fresnel diffraction model, which is much more complex and generally involves detailed computation. Figure 1.13 shows a parallel array of transparent bands of width  $b$ , equally spaced at distance  $b$  from one another. Ideal light transfer would give the crenellated profile shown with dotted lines in the lower image, whereas the actual distribution is the one shown by the continuous curve. The distortion of the intensity profile increases as one moves the mask away from the wafer. Not only is the radiation not uniform in those regions corresponding to the transparent parts of the mask, but a non-negligible intensity is sometimes present in places where no exposure is intended. The more closely the grating interval  $b$  approaches the wavelength of the light being used, the more the intensity profile will deviate from the ideal one. It can be shown that the theoretical resolution limit, i.e., the smallest transferable grating interval, is given by

$$2b_{\min} = 3\sqrt{\lambda\left(s + \frac{1}{2}e\right)}, \quad (1.2)$$

where  $s$  is the separation between the mask and the resist layer and  $e$  is the thickness of the resist layer.

But, even though one can calculate the light intensity during exposure, one cannot necessarily predict the resist profile after development. Indeed, the development process must also be modelled here. The resist contrast makes development a highly nonlinear function of the irradiation level. Hence, a low



**Fig. 1.13.** Light intensity profile on a resist layer, obtained with a mask carrying a parallel array of equally spaced bands of width  $b$ . The *dotted lines* show ideal light transfer and the *continuous curve* is the true light profile on the resist layer

light intensity may have absolutely no effect at the development stage. This modelling can be further complicated by reaction of the resist to irradiation. The transparency of many resists varies with the absorbed intensity. The effects in highly exposed regions are then accentuated compared to those in regions that have been subjected to lower intensities. All these factors can lead to a reduction in the consequences of diffraction effects.

### Contact Lithography ( $s = 0$ )

When the mask is in contact with the resist layer,  $s = 0$  and resolution is maximal. Moreover, resists are media with very high refractive index (of the order of 16) and diffraction effects in the resist are reduced compared to what they would be in air. For a wavelength of 400 nm and a resist layer with thickness  $1\ \mu\text{m}$ , it is easy to obtain resolutions less than the micron. Using a thinner resist layer and shorter wavelength, one can reduce this to  $0.2\ \mu\text{m}$ .

However, obtaining perfect contact is a delicate matter. Both mask and wafer must have perfectly planar surfaces. This is not always possible, because the wafer may carry significant relief produced in earlier processes and not entirely smoothed out by the resist layer. Moreover, the mechanical action required to force the mask against the resist creates debris which may damage both mask and substrate. Likewise, any particles present in the interstice will obstruct perfect contact and reduce the resolution.

Another problem with this technique is alignment. In general, several successive lithography stages are required to produce a single device, and these processes must be very precisely aligned with one another. Alignment

is achieved using marks reproduced on the wafer which are matched to similar marks on the mask. This operation involves displacing one with respect to the other, which means that they cannot be in contact during the alignment procedure itself. The subsequent entry into contact inevitably reduces the accuracy that can be obtained.

It is due to technical problems of this kind, rather than questions of resolution, that contact lithography was eventually abandoned in the mid-1970s, when the critical size of patterns being reproduced was of the order of  $5\mu\text{m}$ .

However, this technique is well-suited to laboratory and R & D work, and some effort has been made to reach dimensions well below the micron. For example, in order to improve the precision with which contact is made, thin flexible masks, also known as conformable masks, have been designed. These bend to fit the wafer surface much more closely and hence yield excellent resolution, below  $0.25\mu\text{m}$  in a resist layer of thickness  $0.5\mu\text{m}$ . Using thinner resist layers and an  $\text{F}_2$  excimer laser which delivers light at  $157\text{nm}$ , features of size  $150\text{nm}$  can be obtained.

### Proximity Lithography ( $s \neq 0$ )

The problems due to surface defects can be solved by introducing a space between mask and resist, although this leads to a rapid degradation in resolution. Indeed, using (1.2), we observe that for any reasonable interval, the minimal period that can be reproduced is given by

$$2b_{\min} \approx 3\sqrt{\lambda s}.$$

Hence for a gap of  $10\mu\text{m}$ , the maximal resolution with a wavelength of  $400\text{nm}$  is of the order of  $3\mu\text{m}$ . Of course, proximity lithography also requires a high degree of planarity in both the mask and the wafer to ensure that the gap between them is constant. In fact, the degree of planarity is generally sufficient to achieve separations as low as  $10\mu\text{m}$ . Below this value, it is difficult to ensure that there are no points of contact between mask and wafer.

Once again, reducing the wavelength improves resolution. It is important to note that, in contrast to lithographic systems by projection, which use optics, wavelength reduction is much easier to implement. Indeed, there are no problems here with absorption or chromatic aberration of the kind that arise in optical systems.

Ease of implementation and low cost make this technique extremely useful for producing items in much smaller numbers than the major microelectronic products, such as memory units or microprocessors. Optoelectronic components which require resolutions of the order of a few microns are still made using this technique. Likewise, microwave and millimeter components on GaAs mainly use proximity lithography and it remains an important tool in the research laboratory.

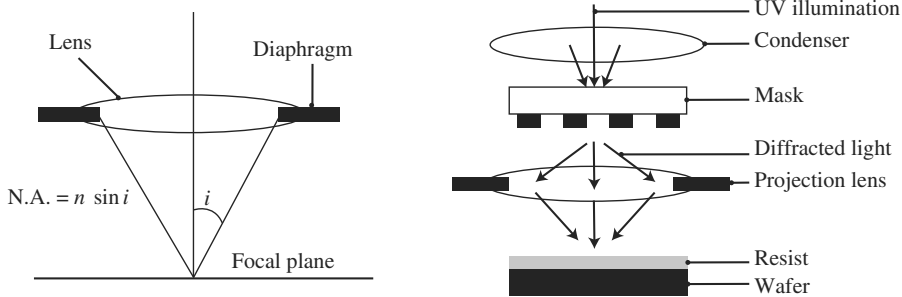
### 1.5.3 Projection Photolithography

The mechanical problems encountered with contact and proximity lithography stimulated the development of projection lithography in the mid-1980s. The mask and wafer are held some distance apart and an optical system is inserted to focus the image of the mask on the wafer. The whole wafer cannot be exposed in a single step, because the field of projection is much smaller than the wafer due to the resolution of the optical system. Two solutions have been devised: either the wafer is moved or the optics are moved.

In the former solution, rather than moving the optical system, which is often cumbersome, it is the wafer and mask that are shifted. This is known as scanning projection printing. The wafer and mask are moved simultaneously and continuously in front of the optical setup. Reflective optics are used, or a combination of reflective and refractive optics, but without magnification, so as to simplify the displacement operation. The advantage with this technique is that one uses only the best zone of the optics and this provides excellent definition. The disadvantage is that there is no reduction of the mask. It must have the same dimensions as the whole pattern to be reproduced on the wafer. Its fabrication thus becomes a delicate matter, and more and more costly as wafer sizes increase and critical dimensions decrease. Alignment accuracy is also hard to improve owing to the mechanical motion, and it is difficult to meet the requirements of size reduction. The increased size of wafers and reduced critical dimensions mean that scanning projection printing is less and less frequently used.

In the alternative approach known as step-and-repeat projection printing, the pattern to be reproduced on the wafer is divided up into identical exposure fields. The mask contains the elementary pattern whose size depends on the resolution of the optical setup. Once the mask has been exposed on the wafer, the latter is shifted along and the operation repeated on the neighbouring field. This continues until a whole row of stepper fields has been exposed. The use of refractive optics makes it possible to reduce the mask size by a factor between 5 and 20, facilitating fabrication and greatly reducing the cost. Obviously, the more the size of the mask is reduced, the bigger will be the pattern on the mask, and since the field of projection is constant, the more repetitions (and hence, the more time) will be required to cover the whole wafer. A compromise must therefore be found between these two factors.

Before each projection, the system must be realigned, and this too increases the time factor, so this approach is slower than the previous technique. However, despite this drawback, step-and-repeat replaced scanning at the beginning of the 1990s, thanks to its superiority in terms of resolution and alignment. The latest steppers now use a hybrid technique known as step-and-scan, in which the mask is scanned by the optics and then the wafer is displaced so as to expose the next field. This allows one to obtain the best of both worlds.



**Fig. 1.14.** Definition of the numerical aperture (N.A.) and optics with mask and wafer

### Resolution

As in proximity lithography, resolution is diffraction limited, i.e., it is determined by diffraction of light at the edges of the opaque regions of the mask. However, in this case, the light is collected only in the far field and one is therefore dealing with Fraunhofer diffraction. This means that geometrical optics, i.e., the theory of plane waves, is applicable. The diffraction pattern is calculated by summing at each point of the image plane the contributions from all wave fronts coming from the diffracting object, taking into account the path length difference in each light path. A very important parameter arises when a lens is used, namely, the aperture. Indeed, diffracted waves make an angle with the optical axis which increases with the order of diffraction. The optical information is contained in all these orders, so the greater the lens aperture, the more complete will be the information gathered, and the better resolved will be the image. We thus define the numerical aperture of a lens by the expression

$$\text{N.A.} = n \sin i ,$$

where  $n$  is the refractive index of the medium in which the waves propagate and  $i$  is the maximum angle at which light is gathered (see Fig. 1.14).

It can be shown that the minimal separation between two objects imaged by a lens is given by the Rayleigh criterion

$$L_{\min} = \frac{0.61\lambda}{\text{N.A.}} ,$$

where  $\lambda$  is the wavelength of the light. This formula only works for waves without spatial coherence. In reality, the light used in photolithography is partially coherent and the numerical prefactor in the Rayleigh formula is closer to 0.5 than 0.61.

However, this discussion presupposes that the lens has no defects and causes no aberration, and also that the light is perfectly homogeneous. In

the real world, these requirements can never be fully satisfied, and we must introduce a factor  $k$  such that

$$L_{\min} = \frac{k\lambda}{\text{N.A.}} .$$

In production processes used in the 1990s, this factor  $k$  was of the order of 0.8. We shall see below how it can be reduced, even to values below the theoretical minimum  $k = 0.61$ . Considerable progress has been made with the optical systems and N.A. has been brought up from 0.28 in the 1980s to N.A. = 0.9 today.

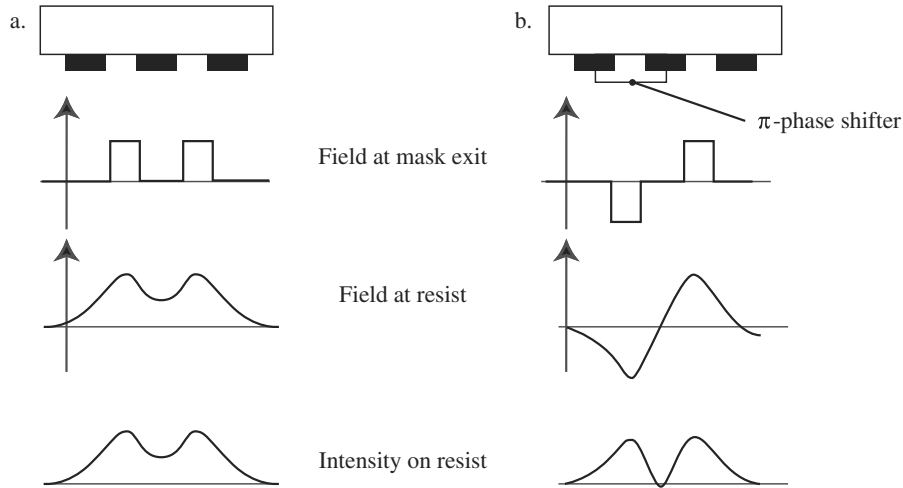
Apart from the technological problems it raises, the increase in the numerical aperture leads to a reduction in the depth of field which is inversely proportional to the square of N.A. A low depth of field exacerbates effects due to defects in the planarity of substrates and the resist thickness. It is generally accepted that there should be a field depth of at least  $0.5\ \mu\text{m}$  in production processes.

Over the years, wavelength reduction has proved an efficient way of enhancing resolution. The 193 nm ArF excimer lasers currently in use have replaced the mercury I line (365 nm) preferred in the 1990s. Production lines using 157 nm F<sub>2</sub> lasers are currently under study and planned for fabrication of devices with minimal dimensions of 65 nm in 2005.

One delicate problem associated with wavelength reduction is absorption in glass. This leads to significant heating effects in the optics, highly complex systems involving a large number of components which must be aligned with great accuracy.

The following methods are used to improve resolution:

- *Off-Axis Illumination.* Diffraction peaks can be shifted using cone-shaped illumination where the rays are highly inclined with respect to the optical axis, thereby increasing the intensity in zones corresponding to the edges of opaque regions.
- *Proximity Optical Correction.* The initial pattern is deformed to account for deformations occurring during projection.
- *Phase-Shift Mask.* Rather than using a mask that contains only transparent and opaque regions, one uses a mask that modulates the amplitude, and also the phase of the light signal. This makes it possible to enhance the contrast of the electric field near dark zones, as shown in Fig. 1.15.
- *Surface Techniques.* The radiation is used to modify the resist surface alone. Problems associated with diffraction in the resist itself and reflection off the substrate are thereby avoided. An example is the silylation process, wherein the irradiated wafer is exposed to a flow of gas containing silicon. The silicon only penetrates to small depths, and only in the irradiated zones. The resist containing silicon is used as a mask for subsequent reactive ion etching. This technique reduces the factor  $k$  and alleviates problems arising from low field depth.



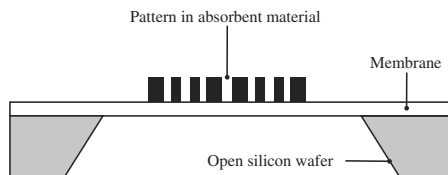
**Fig. 1.15.** Phase-shift mask. (a) Standard mask. Due to diffraction effects, there is a nonzero intensity at a position corresponding to an opaque part of the mask. (b) The  $\pi$ -phase shifter modifies the amplitude of the electric field and thus strengthens the shadow zones

To conclude, optical lithography has made spectacular progress. Features smaller than 100 nm can now be achieved on the production line. A typical step-and-scan machine today has the following characteristics: magnification  $\times 4$ , N.A. = 0.63, field 26 mm  $\times$  33 mm, alignment 45 nm, rate 45 wafer/hr. Such a machine would cost around 10 million euros, as compared with about 0.5 million euros for a DUV (deep ultraviolet) proximity machine.

#### 1.5.4 X-Ray Photolithography

We have seen that the wavelength is the main parameter limiting resolution in optical lithography, but that absorption in glass elements makes it impossible to go below 100 nm. The idea of using extremely short wavelengths in the X-ray region of the spectrum is not a new one. X rays have several invaluable advantages. Apart from the low level of diffraction, they are not sensitive to dust and other organic contaminants, they propagate in straight lines, and they allow a high level of process latitude. But despite these positive aspects and the demonstration that high resolution could be achieved as early as 1975, this technique has never really been adopted in the field of microelectronics. The basic reason, apart from the constant progress in optical lithography which has justified huge capital investment, lies mainly in the problem of masks and sources.

Despite a great deal of effort, no suitable optics has been found to implement X-ray projection lithography. Only proximity lithography is possible. However, there is no X-ray transparent material either, so the mask must be



**Fig. 1.16.** Mask for X-ray lithography

made from a membrane which is thin enough to be transparent to X rays (see Fig. 1.16). The absorbent regions of the mask are made by deposition of a certain thickness of heavy material. Problems of planarity and the fragility of these masks constitute the main obstacle to further development of the technique.

### Choice of Wavelength

The wavelength used is the result of a compromise between absorption in the opaque parts and transparency of the membrane. Thick absorbers with small lateral dimensions can be made, which allow a low limit of the order of 0.5 nm. Fragility of the masks means that one must work with a gap of at least 10  $\mu\text{m}$ . If submicron dimensions are to be achieved, one must therefore use X-rays with wavelengths less than 5 nm.

The resolution limit, if diffraction effects can be neglected, is determined by photoelectrons or Auger electrons emitted during absorption of the photon in the resist. The mean free path of the latter depends on its energy and is of the order of a few tens of nanometers for a 1-nm photon. As it is emitted isotropically, this leads to an effective size for the X-ray photon which can be taken as a tube of radius 10 nm.

Taken together, these considerations lead to a wavelength between 0.5 and 5 nm. In this wavelength range, it is not possible to use reflective optics, owing to surface roughness constraints.

### Sources

X-ray sources are either divergent, e.g., laser-plasma and electron bombardment sources, or parallel, e.g., synchrotron radiation. In the case of a finite point source, a penumbra effect is obtained due to the spatial extension of the source, with a magnification effect depending on the gap, due to beam divergence. The synchrotron provides an ideal source for lithographic purposes, but the complexity in actually putting such a thing into practice has been dissuasive to industrial development.

X-ray proximity lithography has given way to lithography using longer wavelength X-rays (soft X-rays), due to technological difficulties with the mask and the complexity of the source.



### 1.5.5 Extreme UV Lithography

Also known as deep UV lithography, the wavelengths used here are around 13 nm. A more accurate terminology would be soft X-ray lithography. As X-ray masks cannot be used here due to diffraction problems, reflective optics are employed. One further difficulty arises because of the high level of absorption of this radiation in most parts of the apparatus. One must therefore work in vacuum to limit the intensity loss. The reflective optics and masks are made from multilayers, e.g., 40 pairs of Mo/Si at  $\lambda/2$  for radiation at 13 nm gives a reflectivity of 70%. The surface roughness must be very tightly controlled, typically at 0.2 nm/rms, and this equally at small distances to avoid aberrations and at large distances to maintain a high degree of contrast. The low reflectivity of the mirrors means working with rather intense sources. Those envisaged are of laser-plasma type, which are intense but project debris liable to damage the optics. Discharge sources are also under study, but they are not efficient enough.

There remain a good many technological problems to overcome, but EUV lithography is currently the best placed candidate for next-generation lithography. It should replace the 193-nm lines set up in 2000, whilst the critical dimension should go below 50 nm. The first processors to be fabricated by this technique are forecast for 2005.

### 1.5.6 Electron Projection Lithography

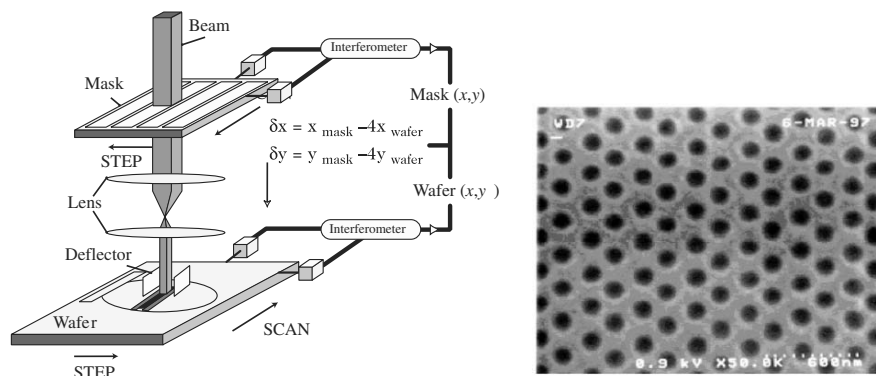
With a wavelength three to four orders of magnitude smaller than photons, electrons are exempt from all diffraction effects in edge regions. We shall investigate the physical limits to resolution with this technique in Sect. 1.5.8, which deals with electron beam lithography. A whole range of electron optics is available to fashion electron beams. The resolution obtained by writing with a focussed electron beam has no equal. However, it suffers from a slowness which rules it out for microelectronic processes. In order to make up for this notorious obstacle, electron projection techniques are under investigation.

Because of the high level of electron absorption in matter, the mask is of stencil type, formed from a silicon wafer. Complementary masks are thus needed to make ring-shaped structures. The electron beam is first broadened into a large parallel beam using a condenser. Two techniques are then possible:

- A mask with ratio 1:1 is placed in close proximity to the wafer and scanned with an electron pencil beam a few millimeters in diameter.
- The mask is irradiated with a broader beam and a lens is used to project the image of the mask onto the wafer (see Fig. 1.17).

With these two methods, the step-and-repeat technique can then be used to expose large areas.

The main difficulties encountered with this technique arise from space charges caused by the strong electron current, which concentrate near the



**Fig. 1.17.** *Left:* Schematic view of the SCALPEL electron projection setup with step-and-repeat. *Right:* Example of SCALPEL production. Holes of diameter 80 nm in a DUV resist of thickness 750 nm. Courtesy of L.R. Harriot, Bell Laboratories: <http://accelconf.web.cern.ch/AccelConf/p99/PAPERS/FRBL1.PDF>

mask and interfere with the beam, and heating of the mask, which leads to distortion. Experimental setups have produced resolutions of 50 nm at a rate of 35–50 wafer/hr.

A novel, massively parallel use of electron beams involves electron microcolumn arrays. This is a technique without masks, in which each microcolumn independently exposes part of the wafer. Another advantage is that the total current is distributed over all the sources and this limits the effect of Coulombic interaction which causes broadening in high-intensity beams. The microcolumns can be a reduction over a few cubic millimeters of a standard electron column. New techniques are also being developed which use microtips in an analogous way to those used in plasma screens. Very dense arrays of electron sources can be obtained in this way. These new sources have very low energy, only a few hundred electronvolts, which allows one to avoid the proximity effects discussed below. However, due to chromatic aberration, one cannot attain probe sizes below 30 nm.

### 1.5.7 Ion Projection Lithography

Because of their high masses compared to electrons, ions constitute a much more efficient means of exposing a resist. They can also be used directly to write or implant a material. Their advantages and disadvantages will be discussed further elsewhere. It is especially for their efficiency and the possibility of high yields that ion projection systems are studied. These systems are very similar to those used for electron projection, except that electrostatic optics are used. The mask is generally of stencil type, although masks using crystalline membranes are an alternative, channelling the ions and thus avoiding beam divergence. Doses are typically of the order of  $1 \mu\text{C}/\text{cm}^2$  compared with several tens of  $\mu\text{C}/\text{cm}^2$  for electrons.

Mask erosion is a serious drawback with this method. It can be limited by using light ions such as helium. Resolutions of the order of 50 nm have been achieved with magnification factor 4. Here again, mask heating and the consequent distortion constitutes the most delicate problem to be resolved before this technique can be used for large scale production.

### 1.5.8 Electron Beam Lithography

This is the preferred method for making masks for optical lithography. Since the work by Ruska in 1930, the considerable progress in generating electron beams and focussing them to make Gaussian probes has led to the rise of electron microscopy. Today, beams of diameter less than 1 nm are produced in a quite routine manner. Electrons have a very small wavelength, viz.,  $\lambda$  (nm) =  $1.22/\sqrt{E(\text{eV})}$ , or 0.04 nm for a 1-keV electron. Edge diffraction effects are therefore negligible. In the 1950s, research was done to assess the possibilities for imprinting resists with fine electron beam pencils. By the 1970s, 10-nm lines had already been achieved.

The basic principle is very simple. The beam scans a substrate covered with an electrosensitive resist, reproducing the required pattern. One thus uses an electron column similar to the one in a scanning electron microscope in which the deflection coils are computer operated. The fundamental difference with all the other techniques mentioned so far is quite clear: this is a sequential process, reproducing the pattern in a stepwise manner, in contrast to global processes like optical lithography. It is therefore much slower and cannot satisfy the requirements of modern microelectronic production. However, its high versatility – because there is no mask, the pattern can be modified at will on the computer – and exceptional resolution make it the instrument of choice for research and development. Phase-shifting masks and the first EUV masks have been made using electron lithography and this equipment is commercialised.

To limit aberration and maintain good focussing, the field scanned by the beam must be restricted. Moreover the digital-to-analogue converters (DAC) which transform the signal from the computer into an excitation in the deflection coils are limited with regard to the number of bits: typically 16 for a DAC of maximum frequency 20 MHz. For a given field, this number of bits fixes the minimum distance between consecutive points on the pattern, thereby defining the pixel. The pixel cannot be bigger in size than the beam without producing dotted lines. A compromise must be found between the probe size, and hence the resolution of the pattern, and the size of the writing field. To give an example, with high resolution so that the probes are 5 nm or less, one uses fields of  $100\ \mu\text{m} \times 100\ \mu\text{m}$ , giving a pixel of 1.5 nm for a 16-bit DAC.

To expose a region bigger than one field, the wafer must be moved, rather as it is in the step-and-repeat system, but this time, the displacement must be controlled with very high accuracy in order to preserve the continuity of the pattern, which generally extends over more than one exposure field of the

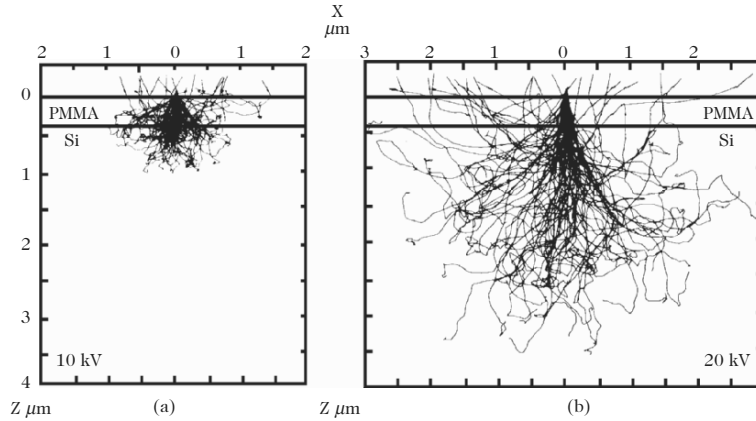
stepper. This is the problem of field stitching. To do this, a laser interferometric device is used to measure the displacement of the substrate holder to an accuracy of the order of the nanometer. Many systems act directly on the scanning, which is corrected to take into account the true position of the wafer, rather than finely adjusting the displacement using piezoelectric motors, for example.

A lithography device can be anything from a simple scanning electron microscope (MEB or STEM), used to expose small areas, to a highly complex machine requiring a temperature-controlled environment, used to expose wafers or 8-inch masks. The characteristics of these machines are expressed in terms of the beam energy, probe size, exposure field, maximum DAC frequency, and accuracy of field stitching.

### Electron–Matter Interaction

Due to their very low mass, electrons are unable to displace atoms by collision. Unlike ions, they cannot directly erode matter. Instead they act rather by modifying the electron structure of the atoms by ionisation. Hence, in organic materials such as polymers, they can break a chemical bond and thereby reduce the chain length of the polymer. Using a weak solvent for this polymer, which only dissolves fragments broken off in this process, one can reveal the regions exposed to the electrons. This is the mechanism applied with PMMA, which is the most commonly used positive resist in electron lithography, due to its high resolution. In the case of negative resists, the effect of the electron beam is rather one of polymerisation. All these polymer-breaking and polymerisation mechanisms involve very low energies, of the order of 5 eV, compared with the beam energies, which are often several tens of keV. One must therefore follow the electron trajectory right down to these low energies in order to analyse the behaviour of the resist under irradiation. There is no closed formula for the energy losses of electrons as they penetrate matter, so one uses numerical simulations of Monte Carlo type.

Figure 1.18 shows the trajectories obtained for point source beams with energies of 10 keV and 20 keV hitting a 0.4- $\mu\text{m}$  layer of PMMA on a silicon substrate. The most striking feature from these simulations is the spreading of the beam due to forward scattering when the electrons penetrate the medium. This broadening increases as the energy of the incident electrons decreases and leads to a loss of resolution relative to the size of the incident beam. One also observes that, far from the point of impact, a large number of trajectories is present, arising mainly from electrons back-scattered by the substrate. The extent of this back-scattering depends on the mass of the material, hence mainly on the mass of the substrate, since the polymers are much lighter. It is responsible for what are known as proximity effects: the dose at a given point depends on the density of pattern features at that point. In other words, it would be difficult to produce a line grating with very small spacing, or worse still, to obtain a small gap between two large pattern features. As can be seen



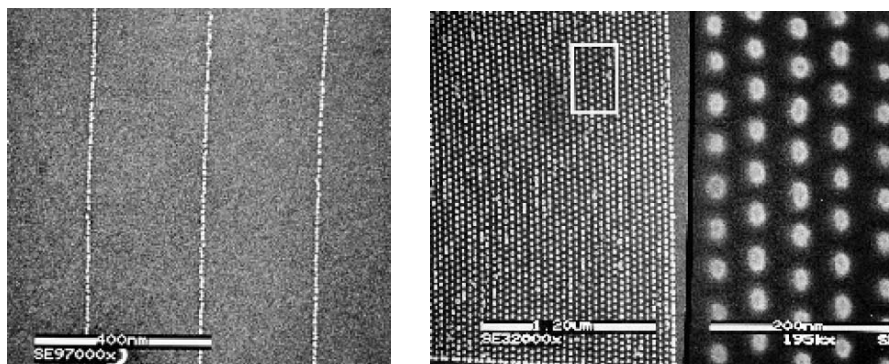
**Fig. 1.18.** Monte Carlo simulations of the trajectories of 10-keV and 20-keV electrons arriving at a point on a silicon substrate coated with  $0.4\ \mu\text{m}$  of PMMA [2]

from Fig. 1.18, the range of back-scattered electrons increases with the beam energy. Calculations can be simplified by representing the energy distribution in the bulk of the material by a double Gaussian function: the first of small width represents the losses due to forward scattering and the second with greater width represents losses due to back-scattering.

### Strategies for Limiting Proximity Effects

As we have seen, these effects do not limit resolution, but they can seriously limit the complexity of a pattern. The following strategies can be used:

- Use of low energies. Unfortunately, this can only be done at the expense of the resolution, since the beam divergence due to forward scattering will then increase. Moreover, owing to chromatic aberration, it is difficult to focus a low energy beam, and this all the more so as the ratio between the energy of acceleration and the energy dispersion of the source is small.
- Use of high-energy electrons. Back-scattered electrons are then diluted over a larger area, which limits their effect on the dose. This is one reason why electron mask machines have progressed from 50 keV to 100 keV over the last few years.
- Calculation of the dose at all points as a function of the overall pattern so as to make local corrections to the dose. Unfortunately, these calculations soon involve divergent computation times as pattern complexity increases. Commercial software is available. Note that such corrections are not always possible because they may require negative doses at certain points!
- Use of resists sensitive only to high energies. Back-scattered electrons lose a great deal of the energy they had in the initial beam, and if the resist



**Fig. 1.19.** Example of resolution limit on PMMA. *Left:* Isolated lines of width 7 nm obtained by gold lift-off. *Right:* SiO<sub>2</sub> dot array with period 40 nm obtained by lift-off followed by etching. Photos C. Vieu

is not sensitive at these lower energies, it will not then suffer exposure by such electrons. Several examples will be given below.

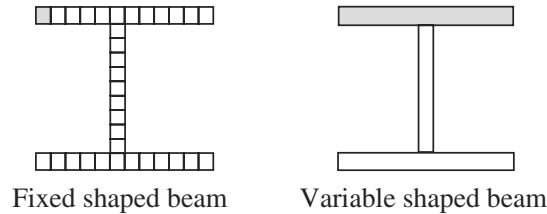
### Limiting Resolution

Organic resists such as PMMA have been used to produce lines finer than 10 nm. The resolution limit with polymers depends on the minimal length of chains that can be broken. Studies have shown that, below a certain length, PMMA chains roll up to form coils with diameters around 5 nm, which would lead to a fundamental resolution limit.

The resolution depends heavily on the dose/development combination. A weak developer gives a high contrast which favours good resolution. The use of ultrasonic waves during development has made it possible to go below the 10-nm threshold using PMMA. The role of these acoustic phonons is to expel polymer residues which stick together by van der Waals forces. This reduces the dose required to open up the lines in the resist. Figure 1.19 shows a grating of isolated lines, in which the interline spacing is much greater than the widths of the lines themselves (5–7 nm). This pattern was generated using PMMA with lift-off, thus demonstrating that the resist was well developed, right down to the substrate. For denser gratings, proximity effects limit obtainable sizes. Periods of 30 nm have been achieved (see Fig. 1.19).

### Inorganic Resists

It was said above that, due to their low mass, electrons are unable to move atoms from their sites. However, at high enough energies, radiolytic effects can cause structural modifications. These are mechanisms whereby electrons transmit their energy to the nuclei of the target atoms, which can then move.



**Fig. 1.20.** Writing strategies for shaped-beam machines

The stoichiometry of certain oxides such as  $\text{WO}_2$  can thus be altered, or they can even be evaporated, under the effects of the beam. Typical energies are of the order of a hundred keV, which are not accessible to back-scattered electrons, but the required doses are several orders of magnitude greater than those used with organic resists. This means that exposure times are very long and the exposure of a usable pattern becomes impossible. In addition, resist thicknesses suitable for this process are very small and this excludes lift-off. Finally, these are materials that generally have low resistance to etching processes. For these reasons, although this type of resist has given remarkable results in terms of resolution – 1-nm  $\text{Al}_2\text{O}_3$  lines have been achieved – it has not been able to produce usable nanostructures.

### Industrial Prospects

Electron lithography is already widely used in industry to fabricate optical masks. Only a few highly specific circuits have been realised in direct write, e.g., the gate level in power transistors for portable telephones. The low yield of this technique limits its use in industry for the lithographic processing of a whole device.

It is conceivable that the current density could be considerably increased in a Gaussian beam in order to improve the write speed. But one must remember the restriction imposed by the speed of the digital-to-analogue converters, and the fact that Coulomb repulsion in the beam can become very strong and prevent correct focussing of the beam.

On the other hand, one might consider using extremely sensitive resists. However, this raises the problem of homogeneity. Indeed, the emission of electrons by the gun, like any discontinuous process, is accompanied by shot noise whose amplitude depends on the square root of the number of electrons emitted. If a resist is highly sensitive and the number of electrons required to expose it becomes very low, noise emissions can attain the level of the required dose. Problems of dose latitude and reproducibility will then arise. It is considered that at least a hundred electrons would be needed to expose one pixel.

Another technique which makes it possible to increase write speeds consists in preshaping the beam. These are referred to as shaped-beam machines. Various setups are illustrated in Fig. 1.20. Most masks for optical lithography

are fabricated with this type of equipment. Of course, the resolution limit is less good than with a Gaussian beam, being typically  $0.20\ \mu\text{m}$ , but the write time can be considerably reduced. It nevertheless remains too long for direct write production to be viable.

### 1.5.9 Focussed Ion Beam (FIB) Lithography

When liquid metal ion sources (LMIS) were developed in the 1970s, it became possible to scan a substrate with a finely focussed ion beam. Indeed, the previously existing gas sources were not bright enough and could not be used in practical situations. Due to their high mass compared with electrons, ions can be used for a wide range of applications: machining, beam-induced deposition, implantation, defect creation, lithography and microscopy.

#### LMI Sources

The idea here consists in wetting a tungsten tip by surface diffusion of a liquid metal. Applying a high voltage between the tip and an extraction electrode, an electric field of the order of  $15\ \text{V/nm}$  is created at the tip apex. This deforms the metal into an even finer point which emits ions. Gallium, which has a melting point around  $30^\circ\text{C}$ , is widely used in LMIS, although liquid alloys are also used (e.g., AuSi, PtB, AuBeSi) to obtain beams of Si, B and Be with the help of a mass separator integrated into the column.

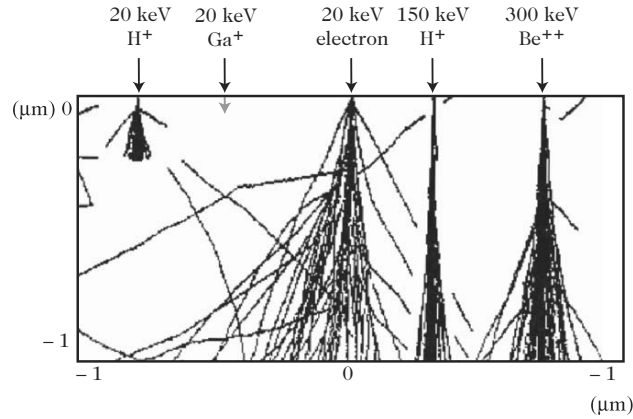
The main drawback with LMI sources is their high energy dispersion, which leads to chromatic aberration and limits the performance of ion optics. It is only very recently that great progress has been made in LMIS design, producing probe sizes below  $10\ \text{nm}$ , with a current density that is compatible with nanofabrication.

The optics used with ion beams involves electrostatic rather than electromagnetic lenses. Indeed, in the latter, the focal point depends on the ratio  $q/M$ , where  $q$  is the charge on the particle and  $M$  its mass. Electromagnetic lenses are therefore of little use with heavy particles. With electrostatic lenses, the focal point is independent of the mass, which is a considerable advantage when using ions, because sources can produce isotopes which then have the same focal point.

#### Ion–Matter Interaction

The main advantage of ions over electrons in lithography, once again stemming from their considerable effective mass and high interaction cross-section, is the low level of scattering, which was precisely the limiting factor in electron lithography. Ion penetration is thus much reduced and occurs in a well defined region, effects that are enhanced as the ionic mass increases. It is then secondary electrons produced by ionic interactions with atoms in the resist





**Fig. 1.21.** Absorption of ion energy for ions with different masses and energies in a 1- $\mu\text{m}$  PMMA layer and comparison with electrons at 20 keV

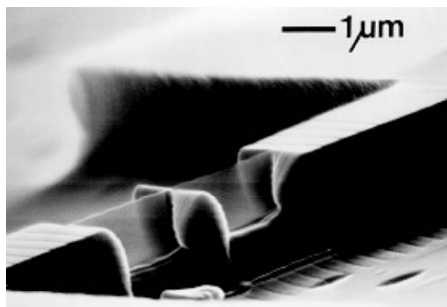
which limit the resolution, whereas in electron lithography, it is the scattering of the much higher energy primary electrons which is the limiting factor. One may estimate that, around an ion trajectory, the resist will be exposed over a radius of the order of 10 nm as the ion goes by.

Figure 1.21 shows how ions lose their energy in matter for a range of different masses and energies. It is quite clear that heavy ions such as Ga are very soon stopped in the resist and are therefore not very well suited to lithography on thick resists, whereas protons penetrate much more deeply. Figure 1.21 also shows the great difference in energy dispersion in the bulk between ions and electrons. This almost complete absence of scattering is a great benefit when ions are used. The very high absorption of ions can lead to critical beam statistics problems. Indeed, if very few ions suffice to expose the resist, statistical fluctuations in the emission can seriously perturb exposure levels. This affects repeatability and can lead to dotted lines, for example.

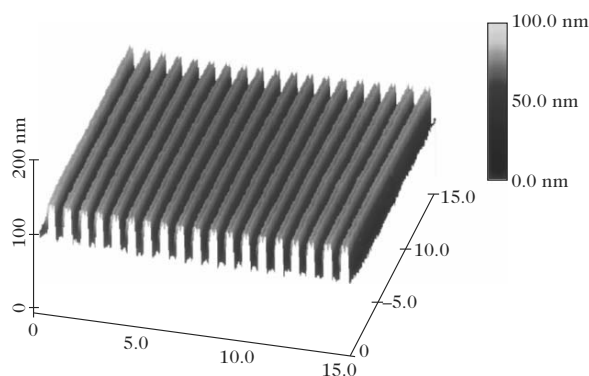
## FIB Applications

*Ion Thinning.* The first applications of FIB used its milling capabilities to thin samples locally in order to prepare them for subsequent TEM (transmission electron microscope) observations in well-controlled regions. Indeed, it is possible to reduce the thickness of a given region by controlled scanning in order to observe its fine structure by TEM. Figure 1.22 shows a chemically etched wall whose width has been reduced using an FIB. The width is now small enough to be transparent to electrons from the scanning microscope used to make this image, which are far less penetrating than those of a TEM.

*Localised Deposition and Reactive Etching.* This function was integrated into commercial machines early on. By introducing a metal-containing gas into the



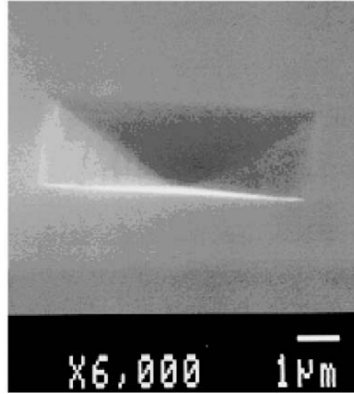
**Fig. 1.22.** Thinning of a GaAs mesa for TEM observation



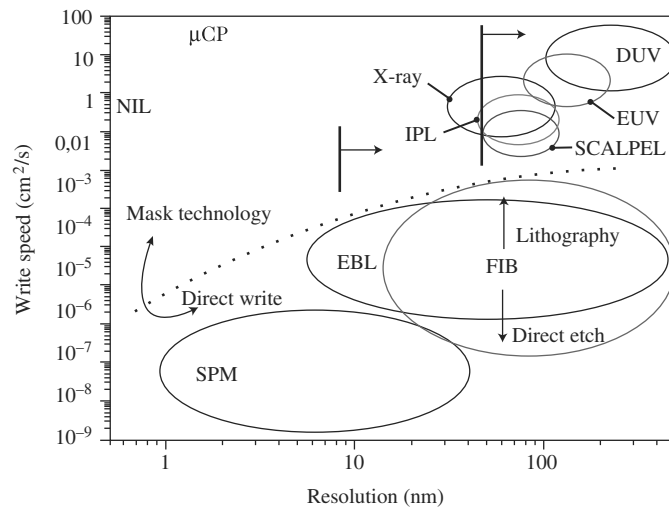
**Fig. 1.23.** AFM image of a series of lines obtained by FIB lithography on  $\text{AlF}_3$ . Note the good verticality of the lines

chamber, the gas molecules can be dissociated by the effect of the ion beam and the metal deposited locally. A classic example uses tungsten chloride gas. Associating this localised deposition technique with etching (without addition of a gas), one can repair or modify optical masks. By injecting a reactive gas, one can also significantly increase the etch rate and suppress redeposition effects by forming volatile chemical components that are then evacuated by pumping.

*Lithography on Inorganic Resist.* These resists are sensitive at high energies. In electron lithography, they can be used to overcome proximity effects. On the other hand, they are poorly sensitive and difficult to use with electrons. Ions are much more efficient and reasonable exposure times can be obtained. Figure 1.23 shows an aluminium fluoride film exposed by FIB. Under the effects of the beam,  $\text{AlF}_3$  decomposes, giving off highly volatile fluorine gas. Only aluminium remains to form the lines. The verticality of the side walls is an indication of the low dispersion of the ions. Note also that the top of the line is hollowed out in the middle. This is due to machining of the upper aluminium layer by the ion beam.



**Fig. 1.24.** Pyramid carved out of a silicon substrate by modulating the dose during scanning by the FIB



**Fig. 1.25.** Panorama of nanolithographic methods on a graph of write speed versus spatial resolution

*Fabrication in Three Dimensions.* In machining mode, it is easy to create 3D structures by judicious variation of the dose during exposure. An example is given in Fig. 1.24, which shows a pyramid carved into a silicon substrate. One application is the in situ fashioning of lenses on vertical-cavity semiconductor lasers.

*Conclusion.* There many applications for focussed ion beams and some manufacturers offer multipurpose machines with an ion column coupled to an electron column for detailed observation. Very recently, it has been shown that focussed ions are extremely effective on magnetic materials. With a low

dose, which means that the process is fast, the magnetisation in a film can be destroyed locally. One can then separate two magnetic domains, without modifying the topology, for these doses are too low to machine the material. This is an important advantage when reading because, if the medium is very flat, the reading head can be very close to the surface, which enhances sensitivity. Lateral diffusion of defects is weak since domains of  $70\text{ nm} \times 70\text{ nm}$  have been demonstrated.

### 1.5.10 Conclusion

We have given here a panorama of different techniques known as far-field techniques that can be used for fabrication. Table 1.1 sums up the main characteristics. For the sake of completeness, one should also consider near-field techniques, in which important progress has been made and which are discussed at length in Chaps. 3–5.

Figure 1.25 graphs the various nanolithographic methods with writing speed on the vertical axis and resolution on the horizontal axis. Naturally,

**Table 1.1.** The main characteristics of far-field techniques in lithography

Technique		Resolution	Use	Comments
Optical lithography	Contact	$0.25\ \mu\text{m}$	Laboratory and R&D	Economical
	Proximity	$2\ \mu\text{m}$	Laboratory and R&D	Economical but low resolution
	Projection	$80\ \text{nm}$	Industry	Spectacular progress
X-ray lithography		$30\ \text{nm}$	Not used at the present time	Development suspended
EUV lithography		$< 50\ \text{nm}$	Industry	Could represent next generation in 2005
Electron lithography	Focussed beam	$1\ \text{nm}$	Laboratory and R&D. Optical mask fabrication	Technique without mask. Best resolution
	Projection	$50\ \text{nm}$	Demonstration	Many difficulties remain
Ion lithography	Focussed beam	$8\ \text{nm}$	Demonstration	Better suited to etching than to lithography
	Projection	$50\ \text{nm}$	Demonstration	Many difficulties remain. Still immature

the sequential writing techniques are located at the bottom of the graph, being the slowest, whilst parallel writing techniques are grouped together at the top of the graph, due to their high writing speeds. The dotted line at the centre of the figure shows the boundary between these two main families of lithographic processes. The speed required for mass production of chips carrying nanodevices is of the order of  $1 \text{ cm}^2/\text{s}$ . It is immediately clear from this graph that there is a problem when we wish to combine such speeds with resolutions close to  $10 \text{ nm}$ .

Lithographic techniques derived from near-field imaging methods are grouped together in the bubble entitled SPM (scanning probe microscopy). They represent the current limit in lithography, achieving atomic scale features in STM mode, but they involve very low writing speeds. Nanoimprint lithography (NIL) and soft lithography (microcontact printing or  $\mu\text{CP}$ ) are very recently developed parallel methods, often classified under the heading of alternative or emergent lithographies, based on polymer moulding techniques. The other processes appearing in the graph correspond to more conventional lithographic methods based on proven tools on the micron scale that have been pushed down to nanometric resolutions.

## References

1. Madou, M.: *Fundamentals of Microfabrication*, CRC Press, 1997
2. Kyser, D.F., and Viswanathan, N.S.: Monte Carlo simulation of spatially distributed beams in electron-beam lithography, *J. Vac. Sci. Technol.* **12** (6), 1305–1308 (1975)

## Growth of Organised Nano-Objects on Prepatterned Surfaces

M. Hanbücken, J. Eymery, and S. Rousset

The invention of near-field microscopy, and in particular, scanning tunneling microscopy [1] and the pioneering work of Don Eigler [2] at IBM Almaden, led to a surge of interest in the manipulation of atoms and molecules. It is the method of choice for investigating individual nanometric-sized objects (which we shall refer to as nano-objects in this chapter). However, the idea of atom-by-atom or molecule-by-molecule manipulation has its own intrinsic limits. For the parallel fabrication of periodically spaced nano-objects of controlled size, self-organised growth on previously structured surfaces (we shall call these prepatterned or prestructured surfaces here) is a novel nanofabrication tool which turns out to be both simple and economical.

Regularity in size is crucial when studying the physical properties of nanostructures (as a function of their size, shape and interactions), because the majority of analysis techniques (e.g., optical and magnetic) are based on averages over large numbers of these objects. Moreover, in many applications such as information storage in magnetic or semiconductor nanostructures, it is essential to assemble nano-objects of similar size in high densities on the surface in order to be able to exploit their individual or collective physical properties.

Organised growth is characterised by the fact that the nano-objects arrange themselves into an array with periodicity dictated by the way the substrate has been prestructured. This is known as the template effect. The substrate is prepatterned either naturally, taking advantage of the surface physics or the first growth stages, or artificially, imposing a given periodicity by lithography and chemical etching. A promising and original alternative is to combine the two aspects, natural and artificial. Hence the distance between nano-objects covers a range from 1 nm to 1  $\mu$ m with a very low size dispersion.

In this chapter, we shall describe the fabrication of nanometric objects with very regular sizes and controlled positions. In all the examples discussed, individual atoms are deposited on prepatterned surfaces which constitute the substrate. This is generally known as the bottom-up approach, as opposed to the top-down approach, which consists in directly etching thin films using

lithographic techniques. The top-down approach is currently too limited to obtain sizes of nanometer order or sufficiently well controlled side walls on fabricated features. Bottom-up methods based on crystal growth can be used to construct objects from the smallest dimensions (one or two atoms) up to larger dimensions (several thousand atoms).

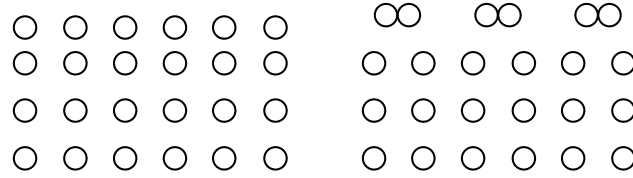
Once the atoms are bound to the surface or to each other, they form the adsorbate. The substrate and adsorbate are often very different from one another chemically speaking, and a given adsorbate does not necessarily form a nano-object on each substrate. In order for this to happen, several conditions must be fulfilled. The physical parameters governing organisation will be described in Sect. 2.1, and their experimental implementation in Sect. 2.2. Then several examples will be given to illustrate these ideas in Sects. 2.3–2.5. The examples in Sect. 2.3 are all based on naturally prepatterned substrate surfaces, wherein a simple preparation of these surfaces in vacuum leads to spontaneous nanostructuring. In Sect. 2.4 an artificial stage involving lithography and chemical etching will be used to prestructure the surface. In Sect. 2.5, we describe an approach combining natural intrinsic structuring of the material with a form of artificial structuring. All these examples illustrate a refinement in the size distribution of the nano-objects, one of the main aims of organised growth as opposed to random growth.

## 2.1 Physical Phenomena in Substrate Prepatterning and Periodic Growth of Adsorbates

In this section, we shall make the distinction between physical phenomena relating to crystal surfaces and the natural prepatterning of such surfaces, discussed in Sects. 2.1.1 and 2.1.2, and nucleation and growth phenomena when adsorbates are deposited on a surface with a view to growing 3D structures, discussed in Sect. 2.1.3. Finally, we shall see in Sect. 2.1.4 that, when the surface is prepatterned by means of artificial techniques, a thermodynamic approach using the chemical potential is better suited to describing island positions.

### 2.1.1 Surface Crystallography: Surface Energy and Surface Stress

A surface is obtained by cutting a crystal along a well-defined crystallographic plane, which fixes its macroscopic orientation. Each surface (also called a face or facet) is labelled by its Miller indices, which specify its normal in a frame of reference that is fixed relative to the unit cell of the crystal. Creating a surface involves cutting bonds between atoms, and the sum of all the forces acting on atoms in the surface is no longer zero. The atoms must shift position in order to reach a new equilibrium state. There are two types of atom displacement: relaxation, which reflects a change in the interplane distance (in general, the



**Fig. 2.1.** *Left:* Cross-sectional view of normal relaxation. *Right:* Cross-sectional view of reconstruction. From [3]

surface plane moves closer to the underlying plane), or surface reconstruction, which induces a change in the atomic structure of the surface.

### Relaxation and Surface Reconstruction

Although the bulk crystal structure of the substrate may be unaffected, alterations appear in the atomic layers closest to the surface. The atomic sites differ from those in the bulk, and two cases are distinguished depending on the direction in which they move. One speaks of normal relaxation when the displacements lead to a reduction in the distance between atomic planes perpendicular to the surface, as shown in Fig. 2.1 (left). Surface reconstruction corresponds to a rearrangement of the atomic sites in the plane of the surface, as shown in Fig. 2.1 (right). A reconstructed surface thus exhibits a different periodicity, which is a multiple of the periodicity in the bulk.

In a reconstruction, the periodicity of the arrangement of surface atoms is modified, and the surface generally exhibits bigger periodicities. The reconstruction phenomenon is very common in semiconductors, but less frequent in metals due to the comparatively weaker angular dependence of interatomic bonds. In general, metallic reconstructions lead to a densification of atoms in the surface.

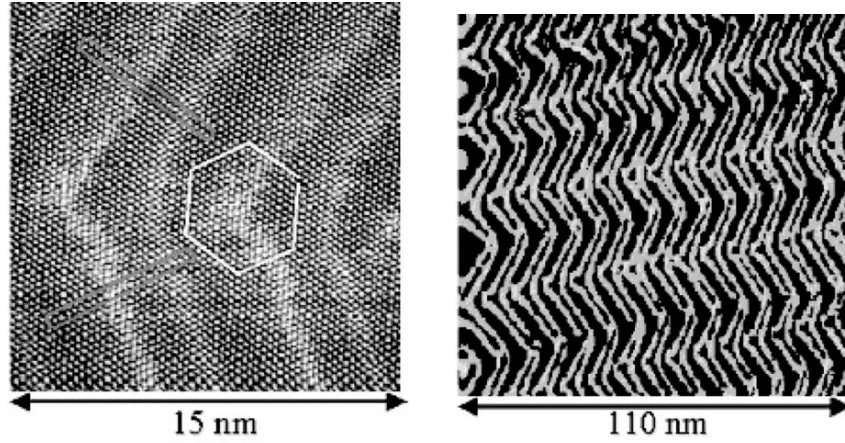
This already constitutes a nanostructuring of the surface, involving periods of a few atomic distances or a few nanometers. However, the reconstructed part of the surface is often very localised, being significantly perturbed by any structural or chemical defects.

### Reconstruction of Gold (111)

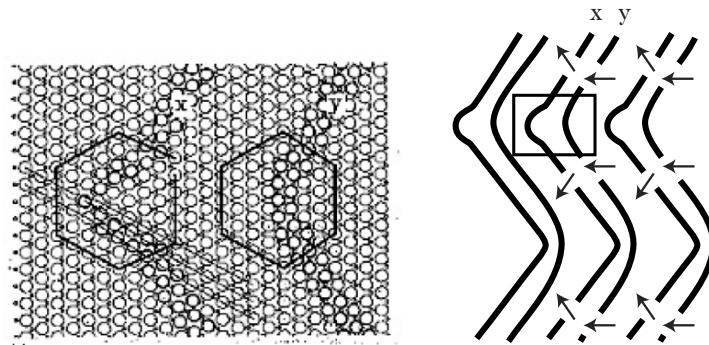
The arrangement of atoms on a gold (111) plane is hexagonal (see section entitled *Crystallography of Surfaces: Vicinal and Dense Surfaces*). This is true locally on the nanometer scale, but on larger scales (see Fig. 2.2), different hexagonal domains are observed, slightly shifted with respect to the others.

On the surface, the atoms are under-coordinated as compared with atoms in the bulk environment, causing them to reorganise and form a reconstruction with surface unit  $22 \times \sqrt{3}$  relative to the  $1 \times 1$  unit of the non-reconstructed surface. This unit cell is a rectangle whose long side represents the direction along which atomic distances have been compressed. Globally, the atoms move together in such a way that, over a





**Fig. 2.2.** *Left:* STM image of the gold surface Au(111) showing the  $22 \times \sqrt{3}$  reconstruction. Two reconstructed domains are shown by *rectangles*. The *hexagon* is a Burger circuit showing the core of a surface dislocation at the join between reconstructed domains. Photo S. Rousset. *Right:* STM image of the herringbone structure of the gold (111) surface. *Brighter lines* forming the zigzags correspond to points on the relief that are some 0.03 nm higher. Photo S. Rousset



**Fig. 2.3.** *Left:* Atomic structure of a bend in the herringbone reconstruction, showing a surface dislocation characterised by its Burger circuit *on the left*, marked by *x*. From [4]. *Right:* Lines of stacking faults organised into zigzags. Bends *x* and *y* have different structures. The *rectangle* indicates the region magnified in the left-hand diagram. From [4]

distance of 22 atoms in the bulk, a 23rd atom is incorporated at the surface. Hence, like a blanket with folds in it, raised lines are created on the surface, corresponding to off-site atoms. These are stacking fault lines, as shown in Fig. 2.3 (right). These faults are clearly visible in STM (scanning tunneling microscopy) images, where they appear as whiter lines, raised by 0.03 nm. In fact, due to the threefold symmetry

of the surface, there are three equivalent reconstruction domains. Two equivalent domains are shown in Fig. 2.3 by the two rectangles oriented at  $120^\circ$  to one another.

The figures also show that the stacking faults form a herringbone structure on the gold (111) surface. This amounts to a periodic structure in the reconstructed domains which can be explained by the self-organisation model developed in Sect. 2.1.2. Each domain involves an intrinsic surface stress which does not have the same orientation. The formation of domains in this zigzag pattern is a way of releasing elastic energy.

The atomic structure of the bends in the zigzags reveals a particular atomic arrangement resembling a dislocation. The presence of such a dislocation can be visualised by plotting a regular hexagon with 10 atoms along each side on the surface (a so-called Burger circuit). The existence of over- or under-coordinated atoms explains why this hexagon is not completely closed, as shown in the figure. Hence this nanostructuring of the surface is known as a dislocation network.

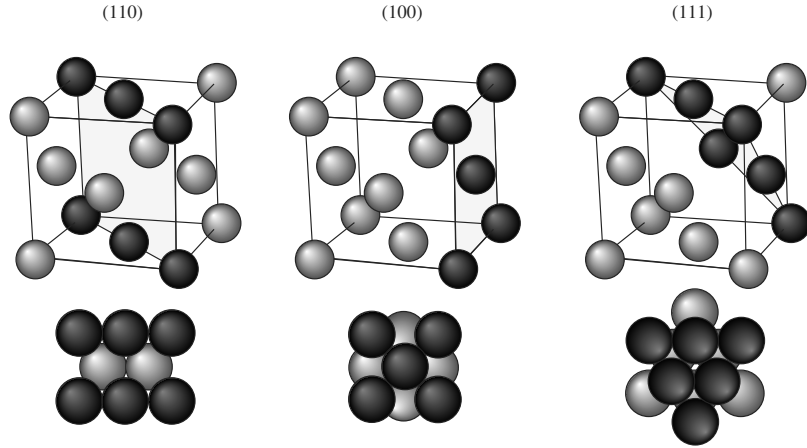
Another, more frequent example of the mismatch between unit cells in the surface layer and in the top plane of the bulk occurs when a metal B is deposited on a different substrate A. One then finds in other systems the same type of domain walls as on gold, but with different and varied arrangements of stacking faults [5].

One simple way to obtain a nanostructured surface with longer-range order is to fabricate a stepped surface. Instead of cleaving a surface along a dense face, one cuts it along a crystallographic plane adjacent to such a face, typically with a misalignment of between  $0$  and  $15^\circ$  with respect to the dense face. The surface produced in this way then exhibits a periodically arranged sequence of terraces separated by steps of atomic height. This atomic stairway is called a vicinal surface. It is the ideal surface on which to grow wires. One may also take advantage of the instability of some of these surfaces which thus evolve towards a factory roof structure, the so-called faceted surface (see examples in Figs. 2.7 and 2.20).

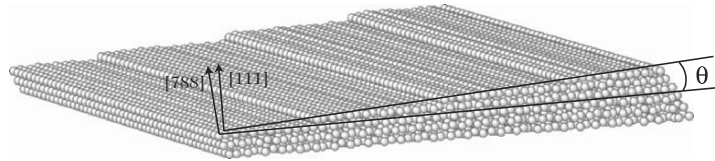
### Crystallography of Surfaces: Vicinal and Dense Surfaces

In a crystal there are dense planes in which atoms are stacked compactly. In a face-centered cubic (fcc) lattice, such as gold, platinum and copper, there are three types of dense face (see Fig. 2.4). Each face is labelled by the direction of the normal vector, whose components are the Miller indices. In Fig. 2.4, the  $(x, y, z)$  axes have been taken as the three sides of the cube representing the unit cell of a face-centered cubic crystal. Face (100) comprises a square arrangement of atoms, while face (111) is hexagonal and is the densest face, and (110) is the most open and the most anisotropic.

A vicinal surface is one with orientation close to that of a dense face. The difference in orientation is somewhere between  $0$  and  $15^\circ$ . Since the positions of surface atoms are imposed in a discrete manner by the crystal lattice, the surface formed by the atoms is no longer a plane but rather a series of steps of height corresponding to the distance between two crystallographic planes. The width of the terraces and hence the density of steps are directly related to the miscut angle  $\theta$  (also known as the vicinal angle) and its direction relative to the crystal lattice (the miscut direction). In the example shown in Fig. 2.5, the crystal cut induces a stairway



**Fig. 2.4.** Dense faces of a face-centered cubic crystal lattice: face (110), face (100), face (111). *Upper:* location of the face in the unit cell of the lattice. *Lower:* view from above, showing the atomic arrangement within the face itself



**Fig. 2.5.** Vicinal surface close to the (111) face, with normal (788). The angle  $\theta$  is the angle by which the vicinal surface is miscut with respect to the dense face. The step density  $n$  is determined by this angle according to  $n = 1/L = (\tan \theta)/h$ , where  $h$  is the step height and  $L$  the width of the terraces

structure on the surface since it creates a simple parallel array of atomic steps. Other cuts relative to the underlying crystal structure can produce several families of steps and hence more sophisticated overlayers, such as a chequered pattern [6].

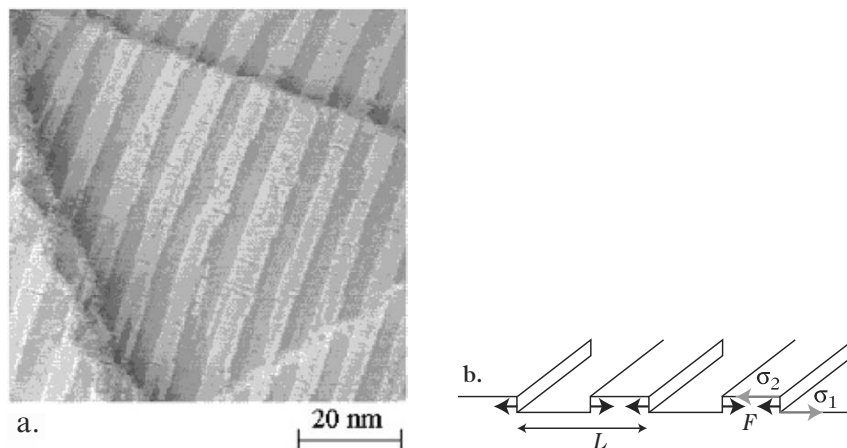
The faceting condition is obtained by minimising the surface free energy, as are the corresponding conditions for relaxation and reconstruction. The free surface energy is defined as the energy needed to create a surface of unit area and arbitrary orientation [7]. It is related to the breaking of chemical bonds when the surface is created. In order to minimise the free surface energy, the surface atoms seek to adopt a different lattice parameter to the one in the bulk. However, since they must adjust to the bulk layer, the surface layer is intrinsically stressed (stretched or compressed). This intrinsic surface stress is analogous to the surface tension in the surface of a liquid, except that there is a fundamental difference between liquids and solids. As liquids are incompressible, when a liquid is deformed, the atoms and molecules move from the bulk towards the surface in such a way as to conserve the density

of atoms on the surface. But when a solid is deformed, the distance between atoms changes and the very nature of the surface also changes. Hence, in a solid, one cannot identify the surface free energy with the intrinsic stress. The surface stress is defined as the energy that must be supplied to deform a surface [8–10]. In the two-phase systems described below, it plays a crucial role in nanostructuring surfaces with long-range order.

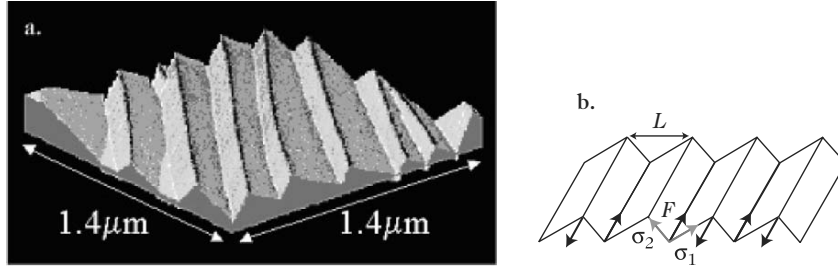
### 2.1.2 Self-Organised Surfaces: Discontinuities in the Surface Stress

Going beyond these crystallographic aspects, long familiar to surface physics, other methods of self-organisation on crystal surfaces have appeared recently, involving the long-range stabilisation of regular structures with periodic features on scales from 1 to 100 nm [10,11]. A spectacular example is the appearance of copper–oxygen stripes when oxygen is adsorbed on a copper surface (see Fig. 2.6a). If a small enough amount of oxygen is adsorbed onto the surface, two phases appear: one is the bare copper surface, and the other is formed from atomic chains of copper and oxygen. In this two-phase system, each phase has a surface energy and hence also an intrinsic surface stress.

At the interface between domains, one finds a discontinuity in the intrinsic surface stress which generates a line of forces between the bare copper and the copper–oxygen domain. These forces allow atomic displacements which



**Fig. 2.6.** (a) STM image of the copper–oxygen striped phase for oxygen coverage of 0.26 monolayers (ML) on the Cu(110) surface (1 monolayer = 1 atomic plane of the substrate, in this case copper). *Dark stripes* formed by copper–oxygen chains alternate with *brighter stripes* of bare copper. From [12] and with the kind permission of P. Zeppenfeld. (b) Periodic alternation (period  $L$ ) of bands of oxygenated copper (stress tensor  $\sigma_2$ ) with bands of bare copper (stress tensor  $\sigma_1$ ) with the presence of forces  $F = \sigma_1 + \sigma_2$  along the boundary between two domains



**Fig. 2.7.** (a) STM image of the faceted structure of a gold (455) vicinal surface. 3D view in which the height scale has been amplified. (The true angle between two successive facets is  $174^\circ$ .) The long-range order has a periodicity of 200 nm. From [16]. (b) Self-organisation model for a faceted surface. The period of the factory-roof morphology is  $L$ . Each facet has an intrinsic surface stress  $\sigma_1$  and  $\sigma_2$ . At the boundary between two facets, there are lines of force  $\mathbf{F} = \sigma_1 + \sigma_2$  which allow relaxation of the elastic energy of the system

release the elastic energy of the system, and elastic deformations propagate into the crystal (long-range term). This elastic relaxation is what causes periodic domains to form in as great a number as possible. For a fixed ratio between the two phases, there is an equilibrium between the energy loss induced by an increase in the density of interphase boundaries and the energy gain which results from elastic interactions between these boundaries. This kind of energy argument can explain the periodicity obtained in Fig. 2.6a, where periodic bands of bare copper alternate with bands of copper covered with oxygen [12].

It should be noted that this is a very general model [9, 10]. It applies whenever the system involves several domains of differing surface stress. This happens, for instance, when an adsorbate does not completely cover a surface [13], but also in clean reconstructed or faceted surfaces. For example, when a surface is faceted as in Fig. 2.7a, each type of facet possesses its own intrinsic surface stress (see Fig. 2.7b). Using similar energy arguments [14], it can be shown that there is a well-defined facet period when the surface reaches its equilibrium structure [15].

### 2.1.3 3D Growth: Energy Criterion and Competition Between Bulk Elastic Energy and Surface Energy

In general, the equilibrium shape of the adsorbate deposited on the surface of the substrate depends on the energy balance between the surface free energies of the two materials (adsorbate and substrate) and that created at their interface [17]. Depending on the relative values of these ingredients, three growth modes are accessible, leading to different morphologies (see section entitled *Growth Modes*):

- In the layer-by-layer mode (also known as the Frank–van der Merwe or FM growth mode), a new layer only begins when the previous one has been completed, and the successive layers tend to spread out.
- In the island growth mode (also known as the Volmer–Weber or VW growth mode), small clusters nucleate directly on the surface of the substrate and the atoms tend to bind to each other rather than to the substrate.
- Another, intermediate growth mode, known as the Stranski–Krastanov or SK growth mode, begins with two-dimensional growth and then continues by three-dimensional growth. In this case, the overall interaction energy between the adsorbed atom and the film varies significantly with the thickness of the deposited film, and this transition can be quite abrupt in many pairs of metallic or semiconductor systems after the deposition of just a few monolayers.<sup>1</sup>

It is the last two growth modes that interest us for the fabrication of nano-objects on a surface. By adjusting the growth conditions, one can thus obtain a certain morphology with some control over the mean roughness and the density of objects. This effect is used to study and exploit physical effects in zero dimensions in components, e.g., optical and electronic properties of semiconductors.

### Growth Modes

Depending on the balance of free energies  $\gamma$  in the adsorbate, the substrate, and the interface between the two, three basic growth modes are possible. Derivations of the thermodynamic quantities playing a role in the growth modes can be found in the review articles [17] and [18]. If

$$\gamma_{\text{substrate}} > \gamma_{\text{adsorbate}} + \gamma_{\text{interface}} ,$$

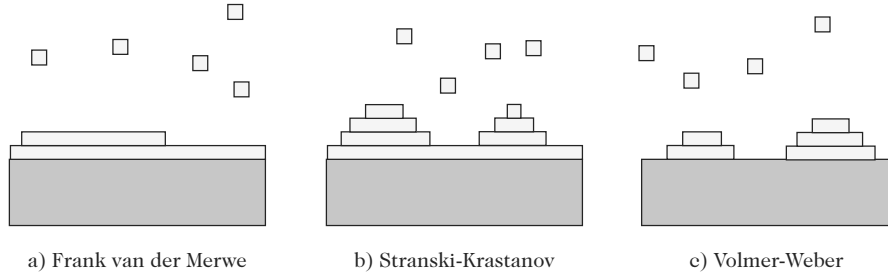
the first layer will tend to wet the substrate. Concerning subsequent growth, there are two possible situations: either the arriving atoms form further layers and the growth is then said to occur layer-by-layer or by the Frank–van der Merwe mode (see Fig. 2.8a); or the free energy of the substrate has already been reduced (and/or the energy due to the lattice mismatch comes into play) and the growth continues in the form of islands on this first layer. This mode is called the Stranski–Krastanov mode (see Fig. 2.8b). If the energy balance is such that

$$\gamma_{\text{substrate}} < \gamma_{\text{adsorbate}} + \gamma_{\text{interface}} ,$$

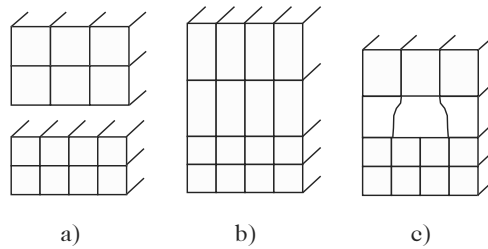
the adsorbate will form 3D islands directly. This mode is called the Volmer–Weber mode (see Fig. 2.8b). It is commonly observed if a reactive material is deposited on an inert substrate, e.g., a transition metal on a noble metal or an oxide.

Very complete tables of surface free energies calculated for gases, alkalis, semiconductors and metals are listed in [19].

<sup>1</sup> The monolayer is the unit of coverage of the sample by the adsorbate. It represents a quantity of atoms adsorbed per unit area equal to that of the substrate surface.



**Fig. 2.8.** The three growth modes



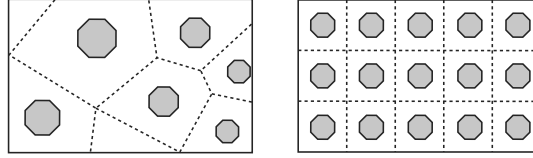
**Fig. 2.9.** (a) Independent layer and substrate. (b) After coherent epitaxy, the deformations are completely accommodated in the plane. (c) After incoherent epitaxy, the lattice mismatch is accommodated by dislocations and a residual stress

In practice, in the Stranski–Krastanov mode, island formation often results from competition between the surface energy which appears on the island faces and the elastic or plastic energy released in the bulk of the island and/or the substrate. This phenomenon, widely used in the case of semiconductors, lies at the heart of the spontaneous self-organisation of quantum dots.

When a solid is in thermodynamic equilibrium, the physical ingredients contributing to the free energy are not only surface and interfacial tensions, but also the parametric mismatch between the elements and the energy of dislocation formation [20]. From the standpoint of the crystal structure, one may distinguish:

- coherent growth of the adsorbate on the substrate,
- totally relaxed growth of the adsorbate,
- partially relaxed growth.

In the first, case, also known as pseudomorphic growth, the adsorbate matches the substrate lattice parameter in the growth plane and as a result, a large amount of elastic energy is stored up in the adsorbed layer when the lattice parameters of the pure materials are very different. The other two cases lead to linear faults known as dislocations, which allow the two lattices to coincide and release energy (see Fig. 2.9).



**Fig. 2.10.** Schematic representation of capture zones for disordered (*left*) and ordered (*right*) nucleation sites. These capture zones may result from a prestructuring of the substrate or a surface reconstruction as in Fig. 2.14

Each of the different cases can be observed experimentally. For the semiconductors InAs/GaAs (III–V) and Ge/Si (IV–IV), the energy cost of dislocations is high and islands are generally coherent with the substrate. This is not the case for the semiconductors CdTe/ZnTe and CdSe/ZnSe (II–VI), where there is first a plastic relaxation (with the appearance of dislocations conserving the planarity of the surface), and then the 3D elastic transition. Using SK growth, one can produce plane surfaces carrying islands whose equilibrium shape can be defined by crystal facets for specific growth conditions. These facets depend on the anisotropy of the surface energy of the islands and the mechanisms whereby elastic energy is released [21]. For example, for SiGe alloys deposited on Si(001), the early stages of deposition produce small objects with  $\{105\}$  facets, while increased amounts of deposited matter lead to bigger objects with  $\{113\}$  facets [22]. When the material is changed, the nature of the facets changes too. For example, one obtains  $\{100\}$  facets for PbSe/PbTe(111) growth [23].

In conclusion, the SK growth mode can be used as a natural way of producing nanoscale objects with relatively well-defined sizes.

To improve the size distribution of adsorbed islands and obtain a regular spacing of islands on the substrate, further parameters need to be considered. Up to now, we have discussed only the equilibrium morphologies, completely disregarding the atomic structure of the substrate surface. Indeed, we have considered a homogeneous surface as far as atomic sites are concerned, and this is not justified in the case of a prepatterned surface. In order to understand this, one must describe growth in atomistic terms, i.e., on the scale of the elementary processes occurring there.

On a plane crystal surface which is chemically homogeneous, all surface atoms are equivalent and growth will be homogeneous. Atoms arriving on this surface can move around, to a greater or lesser extent depending on the temperature, by hopping from one site to another on the surface. During this diffusion process, if two atoms meet, the simplest model assumes that the dimer thereby formed ceases to diffuse: this is then the island nucleation phase. At a later stage in the growth, the number of islands no longer increases, but atoms cluster on existing islands, thus increasing their size. This defines a capture zone (see Fig. 2.10). This growth process on a homogeneous surface leads to a characteristic distance between islands and a mean island size,



both related to the flux of deposited atoms, the temperature of the substrate, and surface defects [22]. However, the island size distribution remains broad, because the quasi-random nucleation/growth sites provide highly fluctuating capture areas for island growth.

In contrast, on a periodic prepatterned surface, the existence of a periodic distribution of favoured nucleation sites, or diffusion barriers which trap deposited atoms, can produce an array of regularly-sized nanostructures in a periodic arrangement across the surface.

#### 2.1.4 Role of the Chemical Potential as Driving Force Behind Adsorbate Growth. Curvature Effect and Elastic Stresses

If the surface is prestructured on a large scale, a more macroscopic description of the growth must be given to explain the way objects are positioned. In doing so, one must take into account curvature effects and stresses in the surface features.

When crystal growth is carried out on a nanostructured surface, the realisation and positioning of objects are once again determined by a minimisation of the total free energy of the system. As we saw above, this energy involves essentially two ingredients. The first is the surface energy of pure bodies and the interface, together with their anisotropies which can cause facets to appear. The second is the elastic energy stored in the bulk of the objects and neighbouring media (substrate, deposited film, etc.). The place where the objects eventually grow depends on the relative values of these two physical parameters.

From a more quantitative point of view, it is useful to consider the surface chemical potential

$$\mu = \left. \frac{\partial}{\partial N} (F + PV) \right|_{T,P},$$

where  $F$  is the surface free energy and  $N$  the number of particles in the system of volume  $V$  and pressure  $P$ . Strictly speaking,  $\mu$  is an equilibrium thermodynamic quantity. Here we are using its extension to a local equilibrium in order to apply it to the case of real growth conditions [24].

It can be shown that the gradient of the chemical potential is a driving force for diffusion on the surface. The atomic surface flux  $j$  is given by the Nernst–Einstein relation

$$j = -\frac{nD}{k_B T} \frac{\partial \mu}{\partial s},$$

where  $n$  is the adatom density,  $D$  is the surface diffusion coefficient, and  $\partial s$  is an infinitesimal length [24].

Along a surface described by a single variable  $x$ , work by Herring (1950) [25] and Mullins (1957) [26] leads to the expressions

$$\mu(x) = \mu_0 + \Omega_0\gamma K(x) + \Omega_0 E_s(x),$$

where  $\mu_0$  is the chemical potential of the plane surface,  $\Omega_0$  is the atomic volume,  $\gamma$  is the surface free energy (which depends on the orientation), and  $K(x)$  is the curvature of the surface (negative for a concave morphology). The function  $E_s(x)$  is the local energy due to the surface stress, which essentially accounts for the tangential component when the surface is free.

This highly simplified model shows that, for the growth of elements identical to those of the substrate:

- the surface curvature term favours adatom diffusion towards the bottom of concave morphological features, such as holes, channels and so on;
- the elastic term favours growth in convex regions, such as ridges, peaks and so on, where there is greater release of elastic energy.

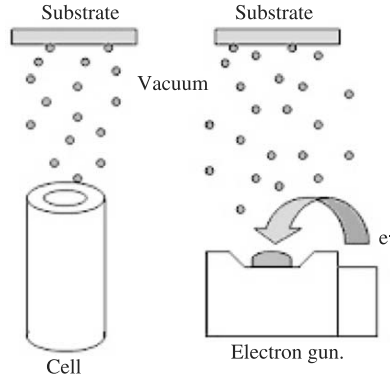
Epitaxial growth of an element differing from the substrate (heteroepitaxy) also depends on these two ingredients. The elastic term must now account for the lattice mismatch between the elements. Hence, the critical thickness at which islands begin to form in the SK transition will be more quickly reached at specific locations in a pattern. One can thus obtain long-range ordering in the dot positions depending directly on the etching interval.

## 2.2 Physical and Chemical Methods for Producing Nano-Objects

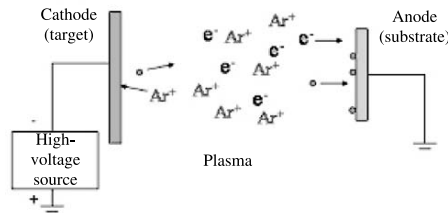
Developments in nanoscience have been widely based on the elaboration of tools designed to deposit thin heterostructure films of semiconductor materials or metallic multilayers (see Figs. 2.11–2.13). This made it possible to move from structures in which just one dimension reached the nanometer scale, viz., the growth axis, to objects possessing two or even three nanometric dimensions.

These fabrication techniques are generally classified into two approaches: the physical approach, in which growth occurs directly from beams of atoms making up the compound; and the chemical approach, involving a chemical reaction which releases those species required for growth.

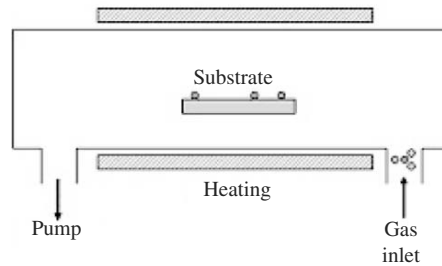
The physical techniques include vacuum evaporation (see Fig. 2.11) in which a material is deposited under secondary vacuum conditions (about  $10^{-6}$  torr, where 1 torr  $\approx$  133 Pa, 1 atm  $\approx$   $10^5$  torr) on a substrate maintained at a controlled temperature, and an extension of this known as molecular beam epitaxy, in which atoms are piled up on a crystalline substrate in such a way as to respect the orientations of the underlying crystal structure, a technique which operates in ultrahigh vacuum conditions (below  $10^{-10}$  torr) and which allows much tighter control of the growth parameters as required for the production of crystalline films.



**Fig. 2.11.** Physical deposition methods operating in vacuum conditions. *Left:* The crucible of the effusion cell is heated by a filament to temperatures as high as 2 000 °C. *Right:* The electron gun heats the material. Another method uses energy from a high-power laser beam (YAG or excimer)



**Fig. 2.12.** Diode sputtering method. The surface is bombarded by a flux of high energy ions (here Ar<sup>+</sup>) obtained from a plasma. The sputtered atoms are generally electrically neutral. Variations involve introducing a reactive element into the plasma, e.g., O<sub>2</sub> or N<sub>2</sub>, or using triode setups or radiofrequencies



**Fig. 2.13.** Chemical vapour deposition (CVD), using a reactor with hot walls. Volatile compounds comprising the material to be deposited may be diluted in a carrier gas. They react on the substrate and the walls. Variations involve using cold walls and heating only the substrate, whereupon one may work at lower pressures, or activating the chemical reaction by means of a plasma

In this technique, the morphological and chemical state of the surface is prepared so as to allow crystal growth plane by plane, with excellent control over the amounts deposited (up to a fraction of a monolayer). The use of several simultaneously evaporated fluxes can produce alloys or compounds with predetermined stoichiometry (e.g., all the binary III–V and II–VI semiconductors). In the case of plane films, one can thereby produce abrupt junctions with well-determined thicknesses, dope semiconductor films by taking advantage of the low levels of residual impurities, and also precisely determine the quantities of matter deposited in 3D islands. Another major advantage of this method arises from the fact that one can use a great many different in situ measurement techniques when the vacuum is good enough, such as diffraction, electron spectroscopy, ellipsometry, or scanning tunneling microscopy. These allow one to study the crystal structures and contamination levels during growth.

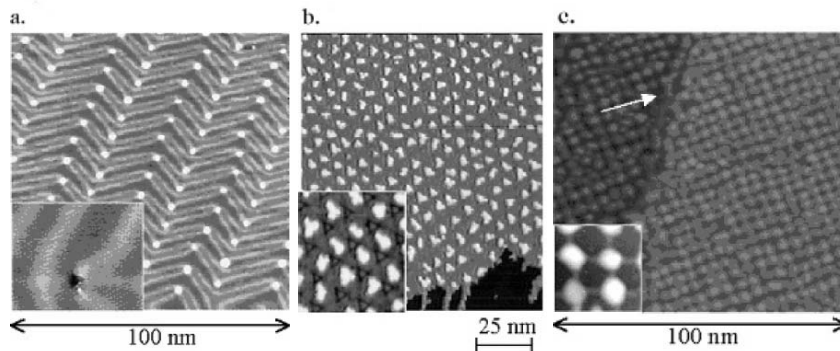
Another physical technique is sputtering, which consists in vaporising the material that will be used to constitute the film (see Fig. 2.12). To do so, a neutral gas (e.g., argon, which does not react with the other species) is ionised by applying either a direct voltage or an ultrahigh frequency field. The positive ions, accelerated by the electric field, then bombard the target (cathode). The transferred energy rips atoms from the target, from whence they are deposited on the substrate.

Chemical techniques for growing films or producing nanostructures use chemical decomposition of a gas or liquid on the surface of the substrate (see Fig. 2.13). With these techniques, it is generally much more difficult to control the amounts of matter incorporated in growth with any accuracy. This category includes all the chemical reactions such as oxidation, nitridation, and so on. The classic example is the oxidation of silicon, so important in microelectronics to produce an insulating layer, where oxygen is supplied in gaseous form. In chemical vapour deposition or vapour phase epitaxy, a gas mixture is used, e.g.,  $\text{SiH}_4$  or  $\text{AsCl}_3$ , to react with the substrate. The gas decomposition can be activated thermally or by applying a plasma. Sometimes organometallic compounds are used (metal–organic chemical vapour deposition or MOCVD), as for example when producing III–V compounds, where an alkyl of a group III metal is made to react with a hydride of a group V metal. The drawback with this method is that it involves the manipulation of highly toxic gases. Finally, in liquid phase epitaxy, the substrate is brought into contact with a dilute solution whose concentration has been chosen to be in thermodynamic equilibrium with the composition of the film one wishes to grow.

## 2.3 Growth of Nano-Objects on a Naturally Prepatterned Surface Using Its Intrinsic Properties

### 2.3.1 Growth of Self-Organised Surfaces

When the surface is self-organised as described in Sect. 2.1.2, the atomic sites on the surface are no longer all identical and one speaks of a heterogeneous substrate. The microscopic processes of nucleation and growth are then considerably altered and lead to periodic growth of islands. In the gold herringbone structure [see the section entitled *Reconstruction of Gold (111)*], the atomic sites on the surface are no longer all equivalent since the bends in the zigzag reconstruction are the scene of surface dislocations, where the coordination number of the atoms is modified [4]. (The coordination number of an atom is the number of nearest neighbours it has. An atom on the surface has lower coordination number than one in the bulk.) These bends constitute favoured nucleation sites for deposited cobalt atoms, because these are able to swap places with gold atoms in the plane of the surface. The driving force behind this insertion is probably due to a reduction of local internal stresses, since the cobalt atoms are ‘smaller’ than gold atoms. Just such a cobalt seed is visible in the inset of Fig. 2.14a. Other cobalt atoms subsequently cluster on these nucleation centers and islands begin to grow. The result is that the surface ends up with periodically arranged rows of islands [28, 29]. The distance between islands in a given row is 7 nm, whilst the distance between rows is of



**Fig. 2.14.** (a) Submonolayer deposit of cobalt on a gold (111) surface. Cobalt islands (*white dots*) form parallel rows. The *inset* shows the nucleation of a cobalt seed (*black speck*) on a preferred site by exchange of cobalt and gold atoms [31]. (b) Growth of silver islands on a platinum surface coated with a silver bilayer. The *inset* shows that the islands are positioned at the centres of the regions bounded by darker stacking faults. With the kind permission of H. Brune. (c) Gold deposited on a Cu(100) surface coated with chemisorbed nitrogen. The *inset* shows a magnified image of four gold islands and the *arrow* indicates step edge defects [32]

the order of 15 nm (see Fig. 2.14a). The size of the nanostructures is regular and varies with the amount of cobalt deposited.

There are other systems for replicating the substrate periodicity during growth. On a platinum surface coated with silver [30], the diffusion of atoms is modified as compared with a homogeneous surface by the presence of stacking faults which act as potential barriers for this diffusion (see Fig. 2.14b). Below a certain temperature (around 110 K), the silver atoms are trapped within a triangular cell and form a cluster at its centre. The atoms deposited on the substrate thus form a hexagonal array of islands (see Fig. 2.14b).

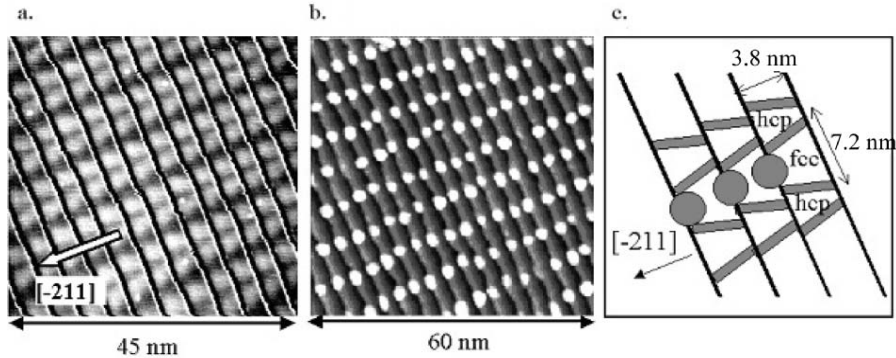
Another example is a square lattice of gold nanostructures elaborated by deposition on a copper (100) substrate coated with nitrogen. For a certain amount of nitrogen chemisorbed at 630 K, the surface is completely paved over by square arrangements of nitrogen atoms measuring 5 nm along the side [13]. By depositing gold on this surface [32], the nitrogen-coated regions act as a mask and the gold atoms congregate on the bare copper regions where they form a square array of regular gold islands (see Fig. 2.14c).

In all these examples, although the atomic processes are very different, the growth is organised, for it replicates the initial nanostructure of the substrate. The islands have a narrow size distribution, close to the state of the art in this field. One novelty of the approach described here for this organised growth is that, not only can one produce nano-objects with very regular sizes, but they are at the same time periodically assembled into an array which also has nanometric dimensions. In this way, one can achieve a high surface density of islands.

### 2.3.2 Uses for Growth on Vicinal Surfaces

On plane surfaces, steps of monatomic height are the most common defects, and they often perturb the organisation of nano-objects on the surface. In some cases, they can favour the growth of nanostructures along step edges, as indicated by the arrow in Fig. 2.14c. Indeed, step edges usually imply different coordination numbers or a specific surface stress, whereupon they become favoured adsorption sites in certain situations.

On a vicinal surface [see the section entitled *Crystallography of Surfaces: Vicinal and Dense Surfaces*], steps are deliberately introduced in a controlled way with regard to density and orientation over the whole macroscopic sample (several square millimeters). On the gold (788) surface miscut by  $3.5^\circ$  with respect to the dense face (111), the stacking faults due to reconstruction are perpendicular to the steps (see Fig. 2.15a). The surface is therefore structured in two perpendicular directions. In one direction, the period is the terrace width of 3.8 nm, whilst in the perpendicular direction, the period is that of the gold reconstruction, namely 7.2 nm. When cobalt atoms are deposited on this surface cooled to 130 K, one obtains an array of cobalt dots with excellent short- and long-range order (see Fig. 2.15b). This is a spectacular example because it combines both the long-range order conserved over the



**Fig. 2.15.** STM images of organised growth of cobalt dots on a gold reconstructed vicinal surface. (a) The Au(788) substrate exhibits steps of monatomic height (0.235 nm) and brighter stacking faults perpendicular to the step edges (0.03 nm). In this image, terrace relief has been subtracted in order to reveal the structure on the terraces more clearly. (b) The same surface with a 0.2 ML of cobalt deposited at 130 K and observed at room temperature [31]. (c) Diagram of three terraces showing stacking faults of the gold surface in *grey* [corresponding to the *whiter lines* in (a)] and cobalt dots [corresponding to the *white discs* in (b)]. The *arrow* indicates the direction  $[-211]$  down the steps of the surface

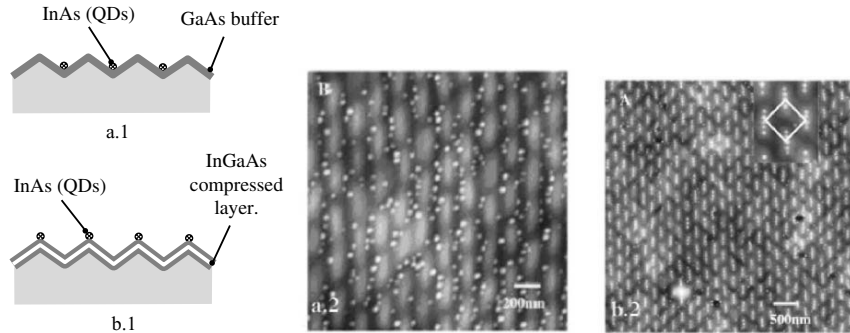
whole microscopic sample and a very narrow size distribution for the cobalt nanostructures.

## 2.4 Growth of Quantum Dots on a Prepatterned Surface by Imposing a Controlled Artificial Pattern

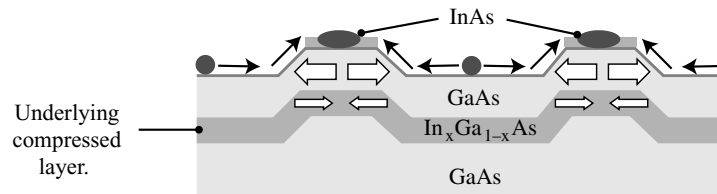
In this section, we describe a surface patterning method which uses both surface morphological features and the influence of local elastic stresses. By selecting suitable epitaxied materials, one can engage upon elastic stress design and engineering.

This method can be illustrated by the InAs/GaAs system ( $a_{\text{InAs}} = 6.058 \text{ \AA}$ ,  $a_{\text{GaAs}} = 5.653 \text{ \AA}$ ), which exhibits Stranski–Krastanov-type growth. The saw-tooth structuring of the GaAs substrate shown in Fig. 2.16 is obtained by optical interferometry and etching. Coherent growth, i.e., monocrystalline and without defects, of InAs islands on the surface of standard GaAs (Fig. 2.16a) occurs at the bottom of concave features. This positioning of the dots suggests that curvature effects are dominant, if we refer to the chemical potential discussed earlier.

However, if a further, coherently compressed layer of  $\text{In}_x\text{Ga}_{1-x}\text{As}$  is first grown in the growth plane with respect to the substrate and then encapsulated with GaAs as shown in Fig. 2.18b, the positions of the InAs dots are completely



**Fig. 2.16.** (a) Quantum dots of InAs deposited on an etched GaAs(001) surface. (1) *Front view* showing the pattern. The buffer layer of GaAs is deposited on the initial etched surface to enhance the early growth stages of the InAs dots. (2) AFM image (atomic force microscopy) of quantum dots deposited on this surface, showing that the dots are arranged in a disordered manner at the bottom of concave morphological features [33]. (b) Quantum dots of InAs deposited on an etched GaAs(001) surface that has been coated with an intermediate compressively stressed layer of InGaAs. (1) *Front view* showing the pattern. Two buffer layers of GaAs have been deposited to enhance structural properties. (2) AFM image. The InAs dots are arranged in an ordered way at the tops of convex morphological features [33]



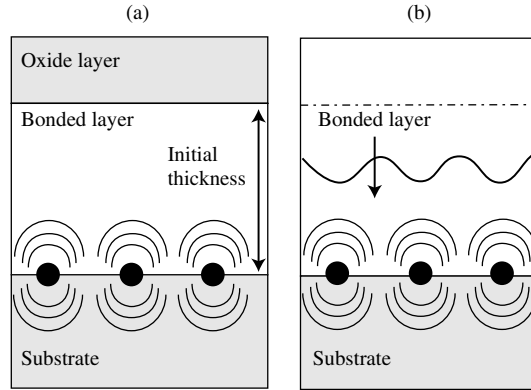
**Fig. 2.17.** Use of epitaxial stresses to localise quantum dots on an etched substrate [34]

reversed, since they will now form at the peaks of the pattern. In this case, it is the stress effect which is so to speak amplified and localised by the surface patterns. Stress relaxation is favoured at the peaks of etched features, as indicated in Fig. 2.17.

The morphology and surface stress can also be nanostructured using an ordered array of buried dislocations which act as a source for stresses extending through the material (see Fig. 2.18). Recent studies have shown that such arrays can be obtained by molecular bonding of monocrystals. This consists in bringing together two clean plane surfaces and annealing to strengthen the adhesion between the materials. For Si wafers, the covalent bonds reform after annealing at high temperature.

The periodicity of the array is controlled by the misorientation angle between the substrate and the bonded film. In this way, one can adjust the spacing interval of the final nanostructure. The problem of controlling the

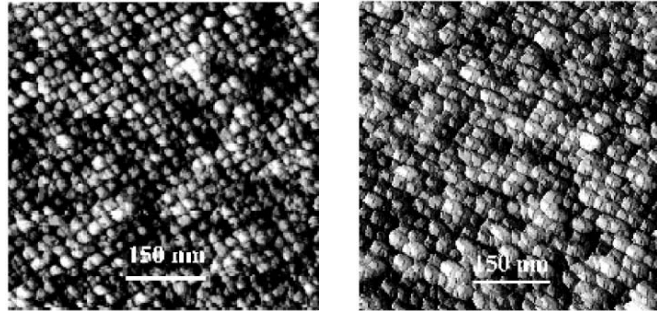




**Fig. 2.18.** (a) Molecular bonding of a thin Si film on a silicon substrate. The surface of the bonded layer is oxidised. Interface dislocations, represented by *black dots*, are the source of elastic stresses shown schematically by *curved lines*. (b) Chemical etching of the bonded layer of the substrate (a) using a solution sensitive to the stresses. The roughness of the surface develops a correlated roughness in the plane with the same period as the dislocations. An example is shown in Fig. 2.19

rotational misorientation angle (twist angle) has been resolved by a method using vernier etches and bonding twin surfaces created by splitting a single wafer [35]. One can thus precisely define the twist of the (001) crystals about the normal whilst almost totally cancelling out other types of misorientation. The annealing operation during molecular bonding causes a highly regular square network of screw dislocations to form in the case of Si(011) (see Fig. 2.19). The level of surface stress depends on how far the dislocations are from the surface. Bondings inducing a periodic deformation of the surface have been used to obtain dots of Ge deposited by molecular beam epitaxy in which the symmetry and interval are directly related to the network of buried dislocations [36]. It has been proposed to carry out chemical etching using a stress-sensitive solution in order to accentuate the surface relief [37]. As can be seen from Fig. 2.19 (left), the almost perfect square periodicity of a buried dislocation network can be transferred to the surface morphology after etching [38].

It has also been shown that surface curvature effects on the nanoscale and inhomogeneous surface stresses completely change the mechanisms of epitaxial growth of Ge on these nanostructured Si surfaces ( $a_{\text{Si}} = 0.5431 \text{ nm}$ ,  $a_{\text{Ge}} = 0.5646 \text{ nm}$ ). Figure 2.19 (right) shows the localisation of matter on top of etched Si islands for a 0.9-nm deposit of Ge at 450°C. This thickness corresponds to about 6.4 monolayers, just beyond the critical thickness of the 2D–3D transition. Standard growth on a plane substrate at the same temperature proceeds by growth of randomly arranged hemispherical islands. This approach can be used directly to organise other organic and inorganic



**Fig. 2.19.** *Left:* STM image of the surface of a bonded and etched substrate ( $0.88^\circ$ , equivalent interval 25 nm, see Fig. 2.18). Roughness rms = 2.46 nm [38]. *Right:* Deposit of 9 Å of Ge at  $450^\circ\text{C}$  on a substrate obtained by chemical etching of a buried dislocation network [38]

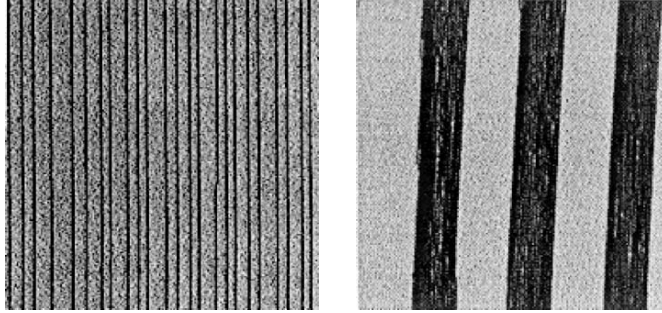
nanometric objects. It is compatible with silicon microelectronics technology and can be integrated over large dimensions.

## 2.5 Growth of Nano-Objects on a Prepatterned Vicinal Surface by Combining Natural and Artificial Patterning

This approach was first developed by T. Ogino and coworkers in Japan [39]. It combines the natural or intrinsic pre patterning of a vicinal surface, in the form of a regular series of steps like the one discussed in Sect. 2.3.2, with artificial patterning, e.g., an array of dimples etched on the surface. This process is detailed here for a vicinal surface of Si(111). We begin by discussing the two types of natural pre patterning known for vicinal Si(111) and then go on to describe an example of a patterned surface which combines the natural and artificial approaches. The growth of gold nano-objects deposited on these surfaces is presented in the next section.

### 2.5.1 Pre patterning the Si(111) Vicinal Surface

An Si(111) vicinal surface transforms naturally into two very different morphologies depending on the cut direction. Vicinal surfaces cut along the direction  $[\bar{1}\bar{1}2]$  rearrange themselves into a pattern of small terraces separated by monatomic steps. Figure 2.20 (left) gives an example of such a surface with a highly regular series of steps. The step height is 0.31 nm with a mean terrace width of 15 nm [40]. If the substrate is cut in the opposite direction, i.e.,  $[11\bar{2}]$ , the rearrangement observed is once again a series of terraces and steps, but this time of much larger dimensions. The steps bunch together to form a facet and in fact these macrosteps are known as step bunches. An

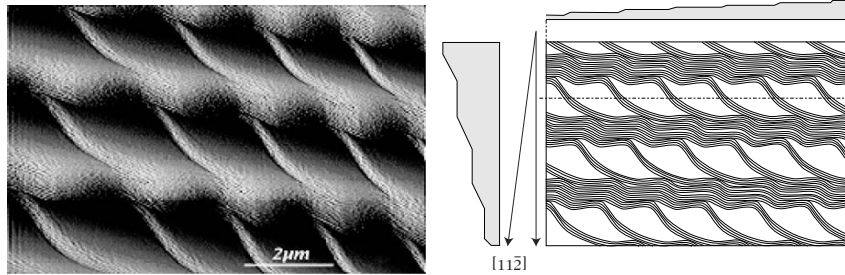


**Fig. 2.20.** *Left:* STM image of an Si(111) vicinal surface, miscut in the direction  $[\bar{1}1\bar{2}]$ . Image size  $340 \times 390 \text{ nm}^2$ . The series of straight monatomic steps was produced by thermal treatment of the vicinal surface [40]. *Right:* STM image of an Si(111) vicinal surface, miscut in the direction  $[\bar{1}1\bar{2}]$ . Image size  $240 \times 240 \text{ nm}^2$ . A prestructure comprising a series of terraces and step bunches with period around 64 nm was obtained by heat-treating the surface

example is shown in Fig. 2.20 (right). The spacing between step bunches has a period of about 64 nm, determined by the intrinsic properties of the substrate, i.e., facet energies, step-step interactions, etc. [41]. Depending on a very carefully established thermal treatment, surfaces can thus be prepared with highly regular structuring for the two miscut directions. They can then be used as templates for subsequent growth processes. The interested reader is referred to the references [42–44].

This kind of spontaneous and intrinsic patterning of Si(111) vicinal surfaces can be modified by adding artificial patterns, e.g., superposing an array of dimples by etching. Substrates etched in this way are heat-treated by the Joule effect (passing a current through the sample) in ultrahigh vacuum. A new rearrangement of the surface is then observed. Due to the presence of the etched features, step bunches are obtained with the formation of facets beside reconstructed terraces. The periodicity is now imposed by the lithographically transferred pattern and no longer depends solely on the intrinsic properties of the substrate. By varying the parameters of the etched pattern (diameter and spacing of the dimples, alignment with respect to the crystallographic axes of the substrate), the morphology of the surface can be continuously modified.

An example is given in Fig. 2.21. On an Si(111) vicinal surface with miscut direction  $[\bar{1}1\bar{2}]$ , which causes step bunching, an array of dimples was etched with a rotation of  $30^\circ$  in the plane of the surface. The STM image shows the characteristic rearrangement into step bunches between terraces. Two types of step bunch are formed. Due to the rotation of the dimple array, some step bunches remain firmly anchored whilst other, smaller ones cross the terraces at an angle. A diamond-shaped pattern is thus created. The repetition of this pattern across the surface modulates the step bunches with concave and convex features (see the example in Fig. 2.21). It has thus been possible to



**Fig. 2.21.** *Left:* STM image of an Si(111) surface prepatterned by a combination of intrinsic patterning and the imposition of artificial patterning by lithography and etching. The surface was subsequently annealed in ultrahigh vacuum conditions at 1000°C for 10 min [45]. *Right:* Diamond-shaped pattern formed on vicinal Si(111). The two types of step bunch are visible. The smaller ones cross terraces at an angle. *Arrows* indicate the misorientation direction and the twist in the plane of the surface

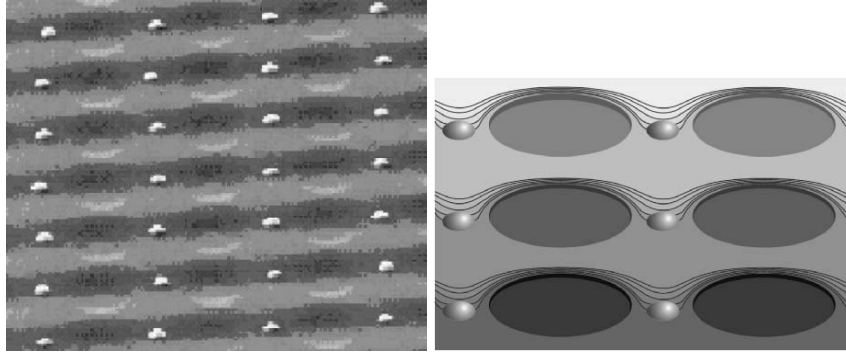
create favoured growth sites on the substrate, corresponding to the discussion in Sect. 2.1.4.

In the next section, we discuss the use of a similar prepatterned silicon surface for the growth of gold nano-objects.

### 2.5.2 Growth of Gold Nano-Objects on Prepatterned Si(111)

In this example, three monolayers of gold were deposited on a surface that had been prepatterned by the method described above [46]. By a suitable thermal treatment, annealing several times at 560°C, the gold adsorbate forms islands at preferred growth sites on the substrate. The annealing temperature was chosen to adjust the diffusion length of gold atoms to the substrate periodicity. The preferred growth sites, with minimum energy, are localised below the step bunch protuberances and between residual etched features. A highly regular arrangement of gold islands with the periodicity of the substrate is thus obtained. The size of the nano-objects can be made as small as desired by changing the number of deposited atoms.

This pre patterning approach is very promising because a change in the parameters of the array etched on the substrate imposes a well-defined pattern. The substrate morphology can be varied continuously and offers a whole range of prepatterned surfaces for the growth of a corresponding range of nano-objects. For example, the density of the nano-objects can be increased by reducing the patterning periodicity. Another strong point of this method is that it can be applied over the whole surface of the sample, whatever its dimensions.



**Fig. 2.22.** *Left:* SEM image of 3 ML of gold deposited on prepatterned Si(111). The deposit was annealed at 560°C for about 20 min. The imaged area corresponds to  $11\ \mu\text{m} \times 11\ \mu\text{m}$  [46]. *Right:* Substrate structured before the gold deposit. The gold islands visible in the left-hand figure are formed in front of the protruding step bunches and between the residual etch features (*grey ellipses*)

## 2.6 Conclusion

The growth of organised nano-objects on prepatterned surfaces is a fast-developing field of activity which rests firmly upon the basic experimental and theoretical principles of surface science. We began by discussing a few important physical ideas which introduce the quantities relevant to organisation phenomena, and then gave a brief review of the physical and chemical techniques of crystal growth. The concept of stress relaxation, whether it be intrinsic surface stresses or epitaxial stresses, plays a fundamental role in these phenomena.

The idea of organised growth on prepatterned surfaces is in fact a rather simple one. The basic technique consists in sending individual atoms onto the substrate and hence growing islands in a natural manner in ultrahigh vacuum conditions. This is both a bottom-up and a massively parallel approach. Nano-objects are fabricated at the limiting size and with high densities. The aim in this chapter was to show that, if the surface used as substrate is prepatterned, the nano-objects then exhibit particularly narrow size distributions, a prerequisite for studying the physical properties of these nano-objects via techniques that average over a large number of objects. When the substrate is prepatterned in a periodic way, island nucleation occurs by replicating the substrate pattern, and this defines capture zones of the same area for growth. As a consequence, the islands all form with the same size and well-defined edges, and they are distributed periodically over the substrate.

The crucial point about organised growth is the way surfaces are prepatterned. Three main categories of preparation used to produce periodic arrangements of nano-objects have been illustrated by experimental examples:

- Using intrinsic properties of the initial surface, such as surface reconstructions, defect networks, or steps on vicinal surfaces. The relaxation of surface stresses underlies the formation of periodic domains on self-organised surfaces.
- Using surfaces directly prepatterned by etching techniques preceded by lithography or the creation of buried dislocation networks. This new field of elastic stress engineering for the organisation of quantum dots, in particular using epitaxial stresses, is developing rapidly and with much success.
- Using surfaces obtained by a combination of artificial patterning and intrinsic patterning. This is certainly a promising channel of investigation for the future, because it guarantees the highest flexibility for the periodicities and sizes of the nano-objects produced.

Several other ideas have been shown to hold promise in the literature, such as the use of ion beams to structure the surface morphology [47], or electrochemistry to create regular pores in materials [48]. All these approaches are designed to address the problem of controlling the positioning and growth of nanometric objects, since this is one of the basic requirements for using such objects in nanotechnology.

## References

1. Binnig, G., Rohrer, H., Gerber, Ch., Weibel, E.: Appl. Phys. Lett. **40**, 178 (1982); Phys. Rev. Lett. **50**, 120 (1983)
2. Eigler, D.M., and Schweizer, E.K.: Nature **344**, 524 (1990)
3. Desjonquères, M.C., and Spanjaard, D.: *Concepts in Surface Physics*, Springer, Berlin (1995)
4. Chambliss, D., Wilson, R., and Chiang, S.: Phys. Rev. Lett. **66**, 1721 (1991)
5. Brune, H.: Surf. Sci. Rep. **31**, 121 (1998)
6. Martrou, D., Eymery, J., and Magnéa, N.: Phys. Rev. Lett. **83**, 2366 (1999)
7. Zangwill, A.: *Physics at Surfaces*, Cambridge University Press (1988)
8. Nozières, Ph.: *Solids Far from Equilibrium*, Chap.1, Cambridge University Press (1991)
9. Ibach, H.: Surf. Sci. Rep. **29**, 193 (1997)
10. Shchukin, V., and Bimberg, D.: Rev. Mod. Phys. **71** (4), 1125 (1999)
11. Alerhand, O., Vanderbilt, D., Meade, R., and Joannopoulos, J.: Phys. Rev. Lett. **61**, 1974 (1988)
12. Kern, K., Niehus, H., Schatz, A., Zeppenfeld, P., Goerge, J., and Comsa, G.: Phys. Rev. Lett. **67**, 855 (1991)
13. Ellmer, H., Repain, V., Rousset, S., Croset, B., Sotto, M., Zeppenfeld, P.: Surf. Sci. **476**, 95 (2001)
14. Marchenko, V.: Sov. Phys. JETP **54**, 605 (1981)
15. Repain, V., Berroir, J.-M., Croset, B., Rousset, S., Garreau, Y., Etagens, V., and Lecoeur, J.: Phys. Rev. Lett. **84**, 5367 (2000)
16. Rousset, S., Pourmir, F., Berroir, J.-M., Klein, J., Lecoeur, J., Hecquet, P., and Salanon, B.: Surf. Sci. **422**, 33 (1999)

17. Venables, J.A., Spiller, G.D.T., and Hanbücken, M.: Rep. Prog. Phys. **47**, 399 (1984)
18. Kern, R., Le Lay, G. and Métois, J.J.: In: *Current Topics in Materials Science*, Vol. 3, ed. by E. Kaldis, North-Holland, Amsterdam (1979) p. 139
19. Sander, D., and Ibach, H.: Surface free energy and surface stress, in: Landolt-Börnstein, *Physics of Covered Solid Surfaces*, ed. by H.P. Bonzel, Springer-Verlag, Berlin (2002)
20. Tinjod, F., Robin, I.-C., André, R., Kheng, K., and Mariette, H.: Journal of Alloys and Compounds **371**, 63 (2004)
21. Tersoff, J., and Tromp, R.M.: Phys. Rev. Lett. **70**, 2782 (1993)
22. Zhang, Z., and Lagally, M. (Eds.): *Morphological Organization in Epitaxial Growth and Removal*, Series on Directions in Condensed Matter Physics, Vol. 14, World Scientific (1998)
23. Raab, A., and Springholz, G.: Appl. Phys. Lett. **77**, 2991 (2000)
24. Villain, J., and Pimpinelli, A.: *Physique de la croissance cristalline*, Collection Aléa-Saclay, Eyrolles (1995)
25. Herring, C.: J. Appl. Phys. **21**, 437 (1950)
26. Mullins, W.W.: J. Appl. Phys. **28**, 333 (1957)
27. Herman, M.A., and Sitter, H.: *Molecular Beam Epitaxy: Fundamentals and Current Status*, Springer-Verlag, Berlin (1996)
28. Voigtländer et al.: Phys. Rev. B **44**, 10354 (1991)
29. Padovani, S., Chado, I., Scheurer, F. and Bucher, J.-P.: Phys. Rev. B **59**, 11887 (1999)
30. Brune, H., Giovannini, M., Bromann, K., and Kern, K.: Nature **394**, 451 (1998)
31. Repain, V., Baudot, G., Ellmer, H., and Rousset, S.: Europhys. Lett. **58**, 730 (2002)
32. Ellmer, H., Repain, V., Sotto, M., and Rousset, S.: Surf. Sci. **511**, 183–189 (2002)
33. Lee, H., Johnson, J.A., Speck, J.S., and Petroff, P.M.: J. Vac. Sci. Technol. B **18**, 2193 (2000)
34. Gerardot, B.D., Subramanian, G., Minvielle, S., Lee, H., Johnson, J.A., Schoenfeld, W.V., Pine, D., Speck, J.S., and Petroff, P.M.: J. Crystal Growth **236**, 647 (2002)
35. Fournel, F., Moriceau, H., Magnéa, N., Eymery, J., Rouviere, J.L., and Rousseau, K.: Appl. Phys. Lett. **80** (5), 793–795 (2002)
36. Leroy, F., Eymery, J., Gentile, P., and Fournel, F.: Appl. Phys. Lett. **80**, 3078 (2002)
37. Wind, R.A., Murtagh, M.J., Mei, F., Wang, Y., Hines, M.A., and Sass, S.L.: Appl. Phys. Lett. **78**, 2205 (2001)
38. Leroy, F., Eymery, J., Gentile, P., and Fournel, F.: Surf. Sci. **545**, 211 (2003)
39. Ogino, T.: Surf. Sci. **386**, 137 (1997)
40. Lin, J.-L., Petrovykh, D.Y., Viernow, J., Men, F.K., Seo, D.J., and Himpfel, F.J.: J. Appl. Phys. **84**, 255 (1998)
41. Men, F.K., Liu, F., Wang, P.J., Chen, C.H., Cheng, D.L., Lin, J.L., and Himpfel, F.J.: Phys. Rev. Lett. **88**, 096105-1 (2002)
42. Himpfel, F.J., Kirakosian, A., Crain, J.N., Lin, J.-L., and Petrovykh, D.Y.: Solid State Commun. **117**, 149 (2001)
43. Li, A., Liu, F., Petrovykh, D.Y., Lin, J.L., Viernow, J., Himpfel, F.J., and Lagally, M.G.: Phys. Rev. Lett. **85**, 5380 (2000)

44. Kirakosian, A., Lin, J.-L., Petrovykh, D.Y., Crain, J.N., and Himpsel, F.J.: *J. Appl. Phys.* **90**, 3286 (2001)
45. Kraus, A., Neddermeyer, H., Wulfhekel, W., Sander, D., Maroutian, T., Dulot, F., Martinez-Gil, A., and Hanbücken, M.: *Appl. Surf. Sci.* **234**, 307 (2004)
46. Homma, Y., Finnie, P., and Ogino, T.: *J. Electron Microscopy* **49**, 225 (2000)
47. Valbusa, U., Boragno, C., and Buatier de Moneot, F.: *J. Phys.: Condens. Matter* **14**, 8153 (2002)
48. Masuda, H., and Fukuda, K.: *Science* **268** (5216), 1466 (1995)



## Scanning Tunneling Microscopy

D. Stiévenard

### 3.1 Introduction

#### 3.1.1 General Principles

The scanning tunneling microscope (STM) was invented by G. Binnig and H. Rohrer in 1982 and they were subsequently awarded the Nobel Prize for Physics in 1986. From an experimental standpoint, the basic idea is as follows: a fine metal tip is brought close to a surface (typically to within one nanometer) and the current flowing between tip and surface is measured when a voltage is applied across the gap. According to classical physics, as there is no contact between the tip and the surface, no current can flow (open circuit). But according to quantum mechanics, if the distance between two electrodes (here, the tip and surface) is small enough, a current can in fact flow across the gap between the tip and the surface. This is the so-called tunnel effect, which has given its name to the microscope based upon it.

The tunnel effect, a purely quantum phenomenon, was first hypothesised in 1927. A particle such as the electron, described by its wave function, has a nonzero probability of penetrating a barrier, although this would be forbidden in classical mechanics. As a consequence, the electron can actually cross a barrier which separates two classically allowed regions. The tunneling probability, i.e., the probability that an electron will pass from one electrode to the other across the barrier, decreases exponentially with the width of the barrier. The tunnel effect can therefore only be observed for narrow barriers, of the order of the nanometer. Theory shows that the current detected is related to the chemical nature of the opposing surfaces, and this on the atomic scale. The microscope is based on a combination of two factors: controlled approach of a metal tip towards a conducting surface, using piezoelectric tubes, and a high-performance anti-vibration system. The piezoelectric tubes have extension coefficients of the order of a few Å/volt and can thus ensure very accurate movements of the tip (bonded onto a piezoelectric ceramic) relative to the fixed surface by applying very low voltages (a few volts).

Binnig and Rohrer demonstrated their invention using a conducting sample and a rather fine conducting tip, which acted as a local probe when brought within a few angstrom units of the surface. With tip–surface voltages of the order of 1 mV to 4 V, tunneling currents of between 0.1 nA and 10 nA were observed. Varying the tip–surface distance established the exponential character of the current as a function of the separation.

STM can therefore be used to observe surfaces with atomic resolution. As we shall see, it can also be used in spectroscopic mode, wherein the tip–surface voltage is varied, to analyse the local electronic structure. Finally, under certain tip–surface interaction conditions, STM allows manipulation of individual objects or even the direct control of local chemistry. It is the only instrument to bring so many benefits: atomic-scale imaging, investigation of electronic structure, and manipulation.

### 3.1.2 General Setup

Figure 3.1 shows the general setup of an STM. A tip is bonded to a piezoelectric tripod allowing motion in the three space directions  $x$ ,  $y$  and  $z$ . The  $x, y$  displacements scan the tip across the surface. The  $z$  axis will reveal the surface topography. A voltage  $V$  is set up between the tip and surface and a current  $I_0$  is chosen. Experimentally, it is the current measured during data acquisition and must be held constant. As we shall see below, the current, the voltage and the tip–sample separation  $d$  are related by

$$I \approx V \exp(-2Kd), \quad (3.1)$$

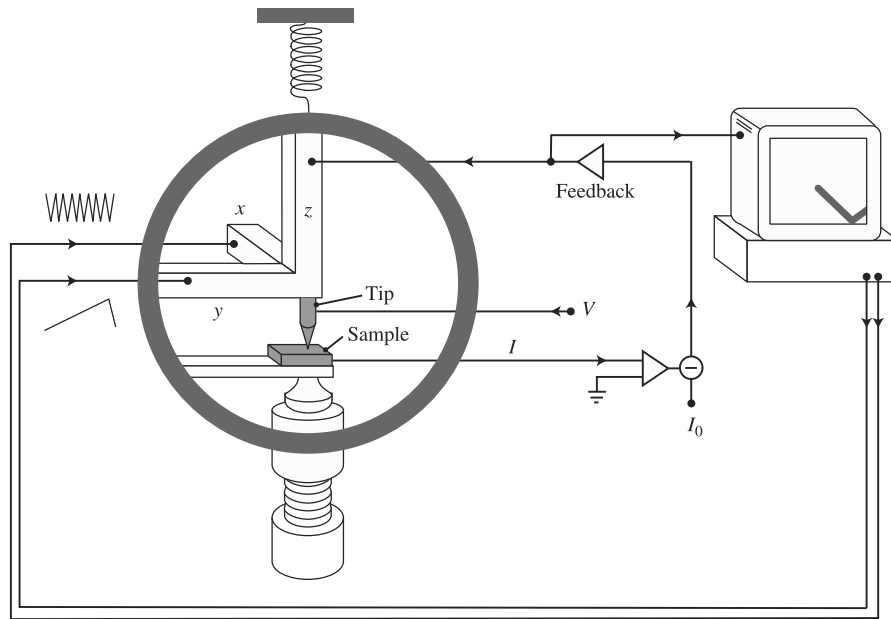
where  $K$  is the wave vector associated with particles in the tunnel barrier, in this case, the vacuum between tip and sample. The tip is brought towards the surface until the distance  $d$  satisfies (3.1). In general,  $d$  is of nanometric order. An  $x, y$  scan is then carried out and the tunneling current  $I$  is measured continuously and compared with the reference value  $I_0$  (constant current operating mode). When  $I$  differs from  $I_0$ , the servo-system instructs the tip to move as appropriate in the  $z$  direction. While varying  $d$  and holding  $I$  constant, the motion of the tip with respect to the surface is recorded. These movements then give the surface topography.

In fact the tunnel current is related to the densities of states of the tip and surface and what is known as the STM topography results from a convolution between purely topographical effects and electronic effects arising from the density of states. The skill of the operator is to deconvolute these two effects in such a way as to produce an accurate interpretation of the STM images. These images are obtained by sending the applied  $x, y$ , and  $z$  voltages to a PC during acquisition. The relief in the  $z$  direction is obtained as a false colour image in which darker to lighter zones are conventionally associated with minimum to maximum regions of the topography.

### 3.1.3 Tip Preparation

For measurements in air, tips are made from a platinum–iridium wire of diameter 100–250  $\mu\text{m}$ . This is an alloy that does not oxidise in air, an essential point, since an oxide layer would insulate the tip and make measurements very difficult. To obtain a fine tip, the simplest method is to cut the wire whilst stretching it, with cutting pincers. One then obtains a stretched wire with tiny barbs or spurs that can be used as nanotips. A certain degree of dexterity is required to achieve a good tip. The advantage with this technique is that, even though it is rather random, it is very quick to implement.

In ultrahigh vacuum, the metal used is tungsten. Tips are prepared by an electrochemical process followed by refinement in ultrahigh vacuum. The tip is thus preshaped by electrolysis. This requires a solution of NaOH (1 mole/l), a molybdenum counterelectrode, a micrometer screw which controls the length of immersed wire, an electronic device which provides a continuous or



**Fig. 3.1.** General setup of an STM. The three piezoelectric tubes displace the tip in the three directions  $x$ ,  $y$ , and  $z$ . A voltage  $V$  is applied between tip and surface. The current  $I$  is measured and compared with a reference value  $I_0$ . The ‘error’ signal is sent into a feedback loop, converted into a voltage, amplified and applied to the  $z$  piezotube. The  $z$  information reaches a PC where it is converted according to a colour scale for display as an image. The  $x$  and  $y$  piezotubes are controlled by the PC in such a way as to scan the sample. If necessary, power amplifiers are used to increase the output voltage of the PC (generally of the order of  $\pm 10$  volts) into the range  $\pm 200$  volts so that an  $xy$  scan of several microns can be achieved

alternating voltage, and a system for detecting tip rupture. There are several stages in the process. First the oxide layer on the wire is removed by applying an alternating voltage of around 20 V peak-to-peak. The wire is then immersed to a length of a few hundred microns and a continuous voltage of around 5 V is applied. This stage generally lasts between 4 and 5 min. The electrolysing current just before rupture is between 150 and 600  $\mu\text{A}$ .

The tip is then transferred to an ultrahigh vacuum. In order to remove the native oxide which forms very quickly with air contact, the tip is heated by electron bombardment. To this end, the tip is positioned in front of a filament and brought to a positive potential with respect to the filament. The voltage is of the order of 600 V for a current of 5 A. This heating cleans the tip, but it also tends to blunt it. Another stage is therefore necessary, namely, ion bombardment. When a tip is used in a medium containing a noble gas, the electrons emitted by the tip can ionise the gas molecules. In this case, as the tip is negatively polarised, it will attract the ions. The latter then move towards the tip and collisions rip off tungsten atoms. This process occurs mainly in places where the electric field is most intense, i.e., close to the tip apex. The sharpest region of the tip thus becomes sharper. The skill of the operator preparing the tip is to control the gas pressure (between  $10^{-6}$  and  $10^{-4}$  torr), the voltage (around 800 V), and the time (between 20 and 60 s), in such a way as to produce as sharp a tip as possible. If the process is allowed to continue for too long, the end of the tip breaks and the process has to be started from the beginning.

## 3.2 Tunnel Current

### 3.2.1 Tunnel Effect Between Tip and Sample

In a barrier of height  $U$ , the wave function  $\psi(z)$  associated with a particle of energy  $E$  less than  $U$  is given by [1]

$$\psi(z) = \psi(0) \exp(-Kz), \quad (3.2)$$

where  $K = \sqrt{2m(U - E)}/\hbar$ , with  $m$  the mass of the particle and  $\psi(0)$  the wave function at the edge of the barrier. This relation shows that the particle state decreases in the positive  $z$  direction. The probability of the particle being inside the barrier is proportional to  $|\psi(0)|^2 \exp(-2Kz)$ , which falls off very quickly with distance. For the tip metal, since the tip-sample polarisation (or bias) is very small in normal operation, the quantity  $U - E$  can be replaced by  $\Phi$ , the work function of the metal.  $\Phi$  is the energy required to transfer an electron from the metal to the vacuum (the last occupied state is the Fermi level  $E_F$ ). The value of  $\Phi$  represents the height of the tunnel barrier between a tip and a metal surface. For metals (tip or sample) or semiconductors (sample) used in STM (W, Pt, Au, Si, etc.),  $\Phi$  is of the order of 5 eV, which means that

$K$  is of the order of  $1 \text{ \AA}^{-1}$ . From (3.1), we see therefore that the current varies by a factor of about ten per angstrom unit. This guarantees a high resolution in  $z$  and it is essential to minimise mechanical perturbation in  $d$ .

### 3.2.2 Tunnel Current: Tersoff–Hamann Theory

The Tersoff–Hamann theory is discussed in detail in [2]. Here we summarise the main features. When the states of the tip and sample are independent, i.e., uncoupled, and for a weak perturbation, i.e., a low voltage between tip and sample and a low temperature, the resulting tunnel effect can be treated as a first order perturbation between independent states (Bardeen approximation [3]), coupled by matrix elements  $M_{\mu\nu}$ , where  $\mu$  and  $\nu$  refer to the two electrodes (here, the tip and sample). Under these conditions, the tunnel current  $I$  is given by

$$I = \frac{2\pi}{\hbar} e^2 V \sum_{\mu,\nu} |M_{\mu\nu}|^2 \delta(E_\mu - E_F) \delta(E_\nu - E_F), \quad (3.3)$$

where  $V$  is the applied voltage, and  $E_\mu$ ,  $E_\nu$  are the energies associated with the wave functions  $\psi_\mu$ ,  $\psi_\nu$  of the electrode (tip and sample) states. The main part of the calculation here is to find the matrix elements. These are given by

$$M_{\mu\nu} = \frac{\hbar}{2m} \int_S dS (\psi_\mu^* \nabla \psi_\nu - \psi_\nu \nabla \psi_\mu^*), \quad (3.4)$$

where  $S$  is an arbitrary surface located within the barrier.

For low  $V$ , a constant potential can be assumed within the barrier and the solution of the Schrödinger equation inside the barrier can be obtained analytically. If one takes the  $s$  wave for the tip states (an analogous calculation can be carried out for the  $d$  and  $p$  waves) and plane waves for the surface states, the calculation of the matrix elements simplifies significantly. The current  $I$  then becomes

$$I \propto \frac{e^2 V}{\hbar} \rho_s(r_0, E_F) \rho_t(E_F), \quad (3.5)$$

where  $\rho_s$  is the density of states of the surface measured at the tip position  $r_0$ , and  $\rho_t$  is the density of states of the tip at energy  $E_F$ . One thus finds that the STM current is directly related to the local density of states (LDOS) of the observed surface.

### 3.2.3 Extending the Tersoff–Hamann Theory

In fact, the low  $V$  case (a few mV) is only applicable to metals and is unrealistic when studying semiconductors, in which case the applied voltage is of the order of a few volts. For semiconductors, a higher voltage is required due to

the band gap. If the voltage is too small, the tunnel effect cannot operate from or to states in the band gap, because there are no states there! In this case, the potential inside the barrier is no longer constant, and what is more, the low-coupling approximation is no longer valid. The Bardeen formalism cannot be applied. A qualitative expression for the current has been proposed by Selloni and coworkers [4], taking into account the trapezoidal shape of the potential in the barrier and calculating the wave functions using the WKB approximation [1]. The effect of the voltage is then expressed through a transmission coefficient  $T(E, V)$  and the current is given by

$$I = \int_0^{eV} \rho_s(r_0, E) \rho_t(r_0, -eV + E) T(E, eV, r_0) dE, \quad (3.6)$$

where the density of states of the sample and the tip are measured at  $r_0$  (the tip position). For negative voltages with respect to the sample,  $eV$  is negative and for positive voltages with respect to the sample,  $eV$  is positive. The transmission coefficient is given by

$$T(E, eV) = \exp\left(-\frac{2z\sqrt{m}}{\hbar} \sqrt{\frac{\Phi_s + \Phi_t}{2} + \frac{eV}{2} - E}\right), \quad (3.7)$$

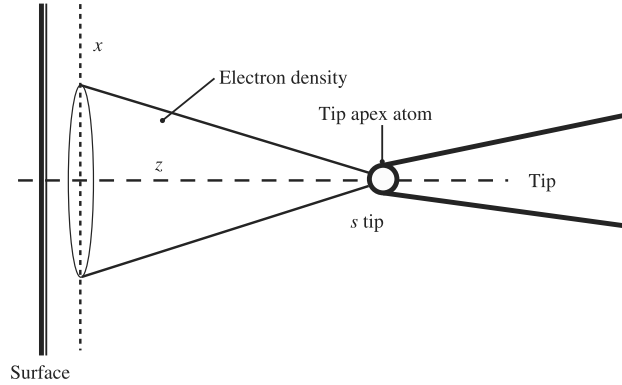
where  $\Phi_s$  and  $\Phi_t$  are the work functions of the sample and tip, respectively. As they are usually close to one another, their half sum is roughly equal to  $\Phi$ .

We thus see that, for constant  $I$ , the path followed by the tip is associated with a rather complicated convolution of the tip and sample densities of states with the transmission coefficient. Indeed, examining the variations of  $T(E, eV)$  a little more closely, we find that, for  $eV < 0$  (negatively polarised surface),  $T(E, eV)$  reaches its maximum for  $E = 0$ , which corresponds to the electrons in the Fermi level of the surface. Likewise, for  $eV > 0$  (positively polarised surface),  $T(E, eV)$  is maximum for  $E = eV$ , which corresponds to the electrons in the Fermi level of the tip. The transmission coefficient is thus always maximum for electrons with energies at the Fermi level of the electrode (tip or sample) which is negatively polarised, i.e., the electrode which emits the electrons. Generally, the width of the electron energy distribution depends on  $\Phi$  and is of the order of 300 meV, with a contribution decreasing typically to 1 eV below the relevant Fermi level.

### 3.2.4 Resolution

The spatial resolution of an STM depends on the nature of the tip and the relevant sample states [5]. A simple approach has been provided by Sacks [6]. To simplify the calculation, one assumes that there is only one atom at the very end of the tip which participates in the tunnel current. One also assumes an  $s$ -type wave function described by

$$|\psi|^2 \propto \frac{\exp(-2Kr)}{r^2}, \quad (3.8)$$



**Fig. 3.2.** *s* wave electronic density of the STM tip apex atom

where  $r = \sqrt{x^2 + z^2}$ , as shown in Fig. 3.2.

Assuming that  $z$  is much bigger than  $x$ , the squared amplitude of the wave function in a plane close to the surface  $S$  can be approximated by a Gaussian function of  $x$ , viz.,

$$|\psi|^2 \approx \frac{\exp(-2Kz)}{z^2} \exp\left(\frac{-Kx^2}{z}\right). \quad (3.9)$$

This shows that the amplitude decreases with the lateral displacement  $x$ . The full width at half maximum of the Gaussian is

$$\Delta x = \sqrt{\frac{2z}{K}}, \quad (3.10)$$

which gives the order of magnitude of the spatial resolution of the STM. With  $K \approx 1 \text{ \AA}^{-1}$  and  $z$  given in angstrom units, the resolution is of the order of  $1.4\sqrt{z} \text{ \AA}$ . In the Tersoff–Hamann theory,  $z = d + R$ , where  $d$  is the tip–sample distance and  $R$  is the radius of curvature of the tip apex. The best resolution is therefore of the order of  $1.4\sqrt{R} \text{ \AA}$ . The practical problem here is to obtain accurate knowledge of  $R$ . In fact the tunneling current is often due to a small protuberance with very small radius of curvature, which explains the images obtained at atomic resolution. The nature of the wave functions associated with the surface states is also relevant. *s* lobes will give less good resolution than *p* or *d* lobes which point less sharply into the gap.

### 3.2.5 Contrast

The contrast of the observed image is something that should be interpreted with great caution because it depends on the measurement conditions (in particular, the tip–sample distance) and the microscopic nature of the surface as given by the lattice parameter  $a$ . Tersoff [7] has given an approximate formula for the contrast  $\Delta z$ :

$$\Delta z \approx \frac{2}{K} \exp \left[ -2z \left( \sqrt{K^2 + \frac{\pi^2}{a^2}} - K \right) \right]. \quad (3.11)$$

For large lattice periods, i.e.,  $a \gg \pi/K$ , the contrast is high and almost independent of the distance. Typically, for  $\Phi = 4 \text{ eV}$ , the contrast tends to  $1.6 \text{ \AA}$  for  $a = 12 \text{ \AA}$ . However, when  $a$  is small compared with  $\pi/K$ , the contrast tends to

$$\Delta z \longrightarrow \exp \left( -\frac{\pi^2 z}{a^2 K} \right). \quad (3.12)$$

It thus decreases exponentially with the tip-sample distance.

### 3.2.6 Measuring the Barrier Height

The key parameter determining the tip-sample distance, apart from the current and voltage, is the height  $\Phi$  of the tunnel barrier. This in turn is determined by the nature of the tip and the sample surface. Moreover, as we shall see,  $\Phi$  is also an essential parameter in STM spectroscopy, for both measurement and interpretation. Depending on the nature of the tip and the small number of atoms located at the tip apex (atoms which may be associated with some contamination or with atoms torn from the surface), the value of the barrier may vary considerably. The tip apex may even evolve under conditions of extreme cleanliness, such as in an ultrahigh vacuum, and it is always affected when STM measurements are made in the air, where pollution and moisture render measurements almost uncontrollable.

From (3.6), when the polarisation is weak, the exponential term varies only slightly with the energy and can be brought outside the integral. The derivative of the current with respect to the tip-sample distance  $z$  is then close to  $I \times 2\sqrt{2m\Phi}/\hbar$ . By analogy, the apparent height  $\Phi$  of the barrier is thus defined by

$$\Phi = \frac{\hbar^2}{8m} \left( \frac{d \ln I}{dz} \right)^2. \quad (3.13)$$

There are three methods for measuring the barrier height. The first consists in direct application of (3.13) to measurements of the current  $I$  as a function of the gap  $z$  between tip and sample, giving  $I(z)$ .

The second method was proposed by Feenstra [8]. It is based on measurements of the conductivity for varying distance  $z$ . It was shown that the apparent height of the barrier is

$$\Phi = \frac{\hbar^2}{8m} \left\{ \frac{d}{dz} \ln \left[ \frac{\sigma'(z_0, V)}{\sigma(z(V), V)} \right] \right\}^2, \quad (3.14)$$

where the conductivity  $\sigma'(z_0, V)$  is given by



$$\sigma'(z_0, V) = I(z_0, V_0)g(z(V), V) \exp \left[ \int_{V_0}^V g(z(E), E) dE \right], \quad (3.15)$$

$\sigma(z(V), V)$  is the measured conductivity, viz.,

$$\sigma(z(V), V) = g(z(V), V)I(z(V), V), \quad (3.16)$$

$V$  is the imposed bias,  $V_0$  is a given bias [ $z_0 = z(V_0)$ ], and  $g = (dI/dV)/(I/V)$ .

The third method involves studying the gap  $z$  as the bias  $V$  varies, so as to obtain  $z(V)$ , when the servo-loop of the scanning tunneling microscope is disabled. As  $V$  increases, the barrier adopts a more and more triangular form, and when  $eV$  exceeds  $\Phi$ , oscillations are detected in  $z$ , associated with the formation of stationary waves between the tip and sample.

In general, the third method gives acceptable values for  $\Phi$ . However, the first two techniques wherein the position of the tip is not fixed relative to the surface can give very different results. In fact, when the tip approaches the surface, several effects occur: there are forces between the tip and sample, and the image potential begins to have a greater influence, deforming the tunnel barrier by addition of a potential  $U$  given by

$$U(z) = \frac{1}{4\pi\epsilon_0} \left[ -\frac{e^2}{4z} - \frac{e^2}{2} \sum_{n=1}^{\infty} \left( \frac{nL}{n^2L^2 - z^2} - \frac{1}{nL} \right) \right], \quad (3.17)$$

where  $L$  is the width of the potential  $U(z)$ .

The origin of the image potential can be understood from simple electrostatic arguments. When an electron is located close to a metal surface, it induces a charge distribution in it. The effect of this distribution is precisely the same as the effect of an image charge of opposite sign placed symmetrically on the other side of the surface. The field in the metal is therefore exactly cancelled, but the shape of the barrier is modified.

The image potential given above diverges to infinity as the surface is approached, something that cannot happen physically. In reality, quantum theory enters the problem and shows that (3.17) is a good approximation to the actual potential if the surface used is an effective surface placed at roughly  $1.5 \text{ \AA}$  from the nuclei of the surface atoms, and if the potential is truncated near the surfaces. Taking into account the effect of the image potential, Chen and Hamers showed that the apparent barrier height can decrease significantly and even tend to zero, agreeing with measurements made on the silicon surface Si(111)- $7 \times 7$  [9].

### 3.2.7 Examples

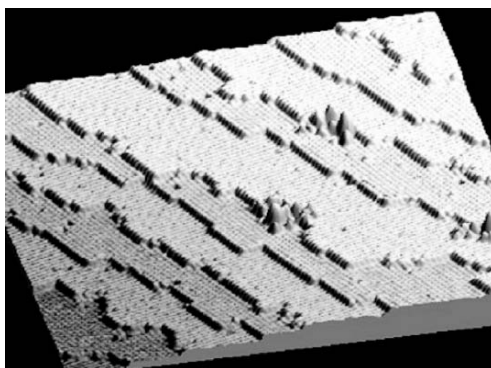
#### Silicon Surface

Figure 3.3 shows an STM image of the  $2 \times 1$  reconstruction of the silicon surface Si(100), observed at 10 K. Clearly visible are the atomic steps, the

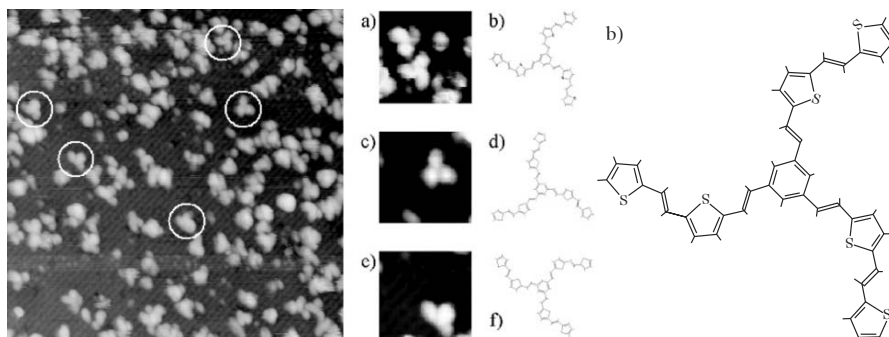
step edges, and the dimer rows on each atomic step. The small piles at the centre of the image correspond to the impact of the STM tip when it has touched the surface.

### Organic Molecule on Silicon

The second example has been chosen to illustrate the capacity of STM to visualise organic nanostructures, in this case organic molecules sublimated in ultrahigh vacuum conditions onto a silicon surface Si(100)- $2 \times 1$ . Another reason for this choice is the increasing role of organic molecules in next-generation microelectronics (see Chap. 13) and the investigation of interactions between single molecules and surfaces. The molecule *n*TV chosen to illustrate this problem has threefold symmetry  $C_{3h}$ . It is made by joining up thienylenevinylene chains *n*TV with  $n = 2$  in this case (see Fig. 3.4b). The three branches are molecular wires with highly delocalised electronic structure. The central benzene ring is an aromatic molecule which also exhibits a high degree of electron delocalisation. In STM imaging, one therefore expects to observe an object with the size and shape of the molecule, whose whole structure should appear as a bright feature. This is indeed what is observed in the STM image of Fig. 3.4, where isolated molecules are visible (encircled in the figure). The size of these features is of nanometric order, as one would expect for a molecule of this kind. What is remarkable is that one observes different conformations of the molecule, as can be seen from Fig. 3.4a–f. In fact, three different configurations of the molecule are observed, associated with rotations of the molecules in the three branches around their single carbon–carbon bond. This example shows that STM can be used to observe individual nanometric objects on a surface.



**Fig. 3.3.** 3D view of a silicon surface Si(100)- $2 \times 1$  observed at 10 K



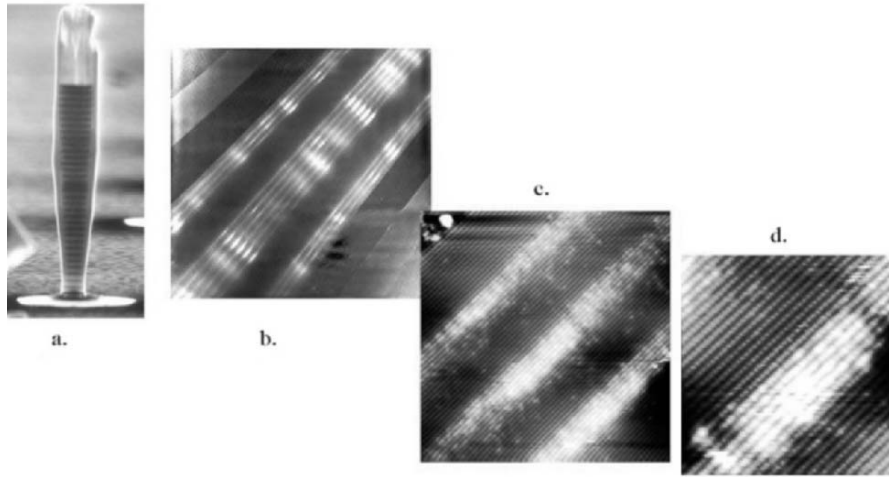
**Fig. 3.4.** Observation of different configurations of an isolated organic molecule on a silicon surface Si(100)- $2 \times 1$

### InAs Quantum Dots in GaAs

The third example concerns the observation of InAs quantum dots (QD) in GaAs (Fig. 3.5). These QDs are obtained by self-organised growth and have unusual properties for the solid state. Of nanometric dimensions, they are the scene of strong electron confinement, which discretises the energy levels to produce 0D systems. As a consequence, the optical properties of QDs are quite remarkable (quasi-monochromatic emission line at low temperatures). This gives them great potential for the fabrication of optoelectronic components (active centers in  $1.3 \mu\text{m}$  lasers) and also in more fundamental research (single-photon emission, quantum cryptography).

Figure 3.5a is a photograph of a micropillar with mean diameter  $1 \mu\text{m}$ , obtained by scanning electron microscope (SEM). The cavity containing the QDs is at the centre of the column and the dark lines on either side are Bragg mirrors which focus the light in the active zone where the quantum dots are located.

The sample has been observed by STM on the cleaved (110) surface parallel to the growth direction [001]. It contains 12 planes of QDs in three groups of three, six and three planes, respectively (Fig. 3.5b). The interplane distance is 13 nm and the groups of planes are separated by a 70-nm layer of GaAs. Each plane of QDs is obtained by depositing 0.55 nm of InAs (1.8 monolayers) in one second at a temperature of  $520^\circ\text{C}$ . Immediately following the deposit, the plane of QDs is buried under a layer of GaAs. In order to pick out the structure easily, successive layers of GaAs and AlAs are used as markers on either side of the QD planes. Figures 3.5c and d show the QDs at atomic resolution. One can then study the size of the QDs (width 15–30 nm, height 3–6 nm) and their interface roughness. The bright appearance of the QDs arises here due to bulging of the surface caused by its relaxation after cleavage.



**Fig. 3.5.** STM images of planes of InAs quantum dots in GaAs

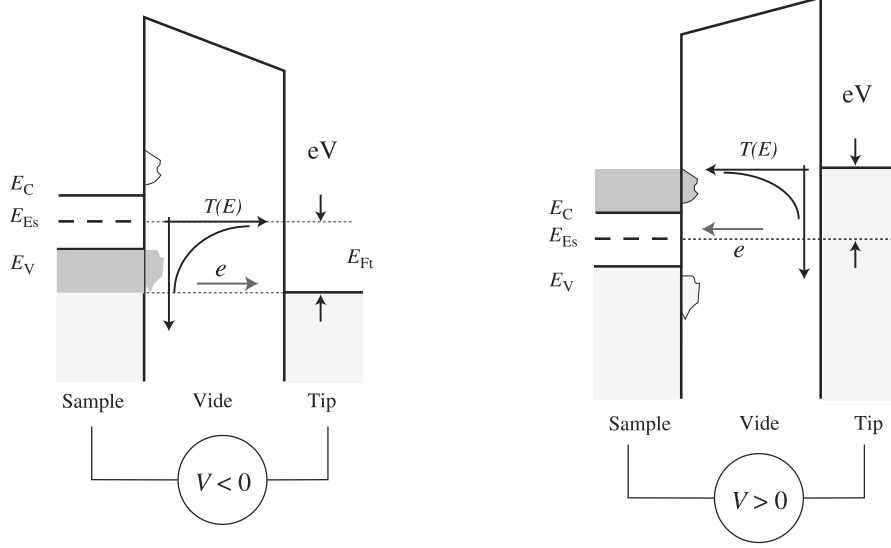
### 3.3 STM Spectroscopy

The second operating mode of the STM is the spectroscopic mode. In this case, the tip is held fixed above the sample surface, the servo-loop is switched on, and a measurement  $I(V)$  is recorded. This produces a local spectroscopic analysis. By varying  $V$ , one analyses the electronic structure of the surface at different energies.

#### 3.3.1 Elastic Current

We begin by considering the case of an elastic current: the electrons do not interact with the surface or the observed nanostructure, whence their energy remains constant. There is no coupling between the electrons and the structure under investigation (weak electron–phonon coupling).

In metals, a low voltage is used (a few hundred meV) and it can be shown that  $T(V) \sim aV$ , to third order and for voltages less than 3 V, which is almost always the case. In this situation, using (3.6), it follows that the derivative of the current with respect to the voltage gives  $\rho_s(E)$ . For semiconductors, due to the band gap, voltages of a few volts are required, typically  $\pm(1-3)$  V, and one must take into account the variation of the transmission probability  $T(E, eV)$  as a function of the voltage. However, Feenstra has shown [8, 10] that a good approximation to the logarithmic derivative of the current with respect to the voltage (the normalised conductance) is independent of the transmission coefficient  $T(E, eV)$ . It is given by



**Fig. 3.6.** Transmission coefficient and  $I(V)$  spectroscopy

$$\frac{dI/dV}{I/V} \approx \frac{\rho_s(E)}{(1/eV) \int_0^{eV} \rho_s(E) dE}. \quad (3.18)$$

This relation shows that one can measure the density of states locally by measuring the function  $I(V)$ . In metals, low voltages are used (plus or minus a few 100 mV). Figure 3.6 shows the band diagrams and probed states as a function of the bias.

For semiconductors that have band gaps without surface states [e.g., the GaAs(110) surface is naturally passivated with a band gap of 2.5 eV], it is difficult to measure the band gap and in particular the transition to the band edge level. Feenstra has suggested bringing the tip towards the surface during the  $I(V)$  measurement, with a maximum gap when  $V = 0$ . (As the voltage is ramped up from  $-V$  to  $+V$ , the tip-sample distance decreases until  $V = 0$  and then returns to its original value.) The current varies exponentially with the distance and this should significantly increase the sensitivity of the measurement.

According to (3.7), the transmission coefficient  $T$  is maximum for

$$T(V) = \exp\left(-\frac{2z\sqrt{m}}{\hbar} \sqrt{\Phi - \frac{|V|}{2}}\right). \quad (3.19)$$

$T$  depends on the polarisation of the junction. Now this polarisation varies during the experiment, but the variation can be compensated by that associated with the variation in  $z$ , in such a way as to keep  $T$  constant during the measurement of  $I(V)$ . One must therefore follow the height curve  $z(V)$  given

by

$$z(V) \times \sqrt{\left(1 - \frac{|V|}{2\Phi}\right)} = z_0, \quad (3.20)$$

where  $z_0$  is a constant equal to the tip-sample distance at zero bias. For biases less than  $\Phi$ , the curve  $z(V)$  becomes [8]

$$z(V) = z_0 \left(1 + \frac{|V|}{4\Phi}\right). \quad (3.21)$$

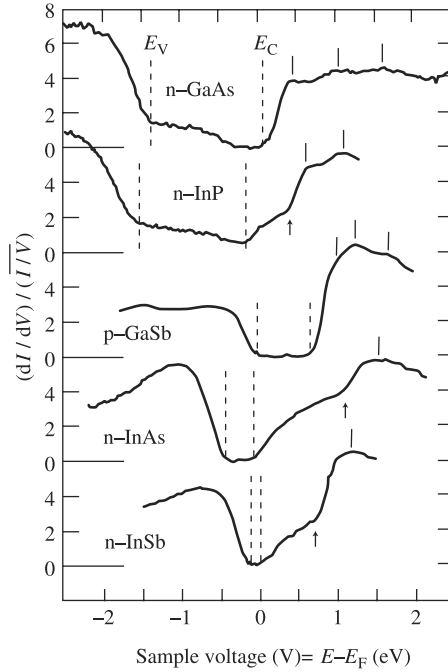
$z(V)$  thus varies linearly with the voltage. The order of magnitude of the slope is  $0.6 \text{ \AA}/\text{V}$  with  $z_0 = 1 \text{ nm}$  and  $\Phi = 4 \text{ eV}$ . This slope depends on  $\Phi$  and can therefore become steeper for small barrier heights.

### 3.3.2 Measuring the Band Gaps of III–V Semiconductors

A good illustration of STM spectroscopy is provided by Feenstra's measurements of semiconductor band gaps, carried out for a whole series of III–V semiconductors (see Fig. 3.7, in which the curves are based on measurements repeated hundreds of times at different points of the surface). The maximums of the valence band and the edge  $T$  of the conduction band are indicated by dashed lines. The energies corresponding to the band edges are determined by the intersections of the horizontal axis with the tangents to the density of states curve in the region where the latter passes from zero to a nonzero value. Some materials have wide band gaps, e.g., GaAs with a gap of 1.5 eV, and others have narrow band gaps, e.g., InAs with a gap of about 0.3 eV. The vertical arrows (for InP, GaSb, InAs, and InSb) correspond to detection of the minimum  $L$  of the conduction band which, for a given voltage (and hence for a given energy), brings about an increase in the density of states of the conduction band. For GaAs, the point  $L$  is not clearly visible, being masked by a surface state. The differences measured between points  $T$  and  $L$  are in good agreement with the calculated band structures for the materials considered. Finally, the small resonances marked by short vertical dashes are associated with surface states appearing at the highest energies for the (110) surface. The number of observed peaks has not yet been completely explained, however. The error in the peak positions is of the order of 0.03 eV.

### 3.3.3 Spectroscopy of Individual Quantum Dots

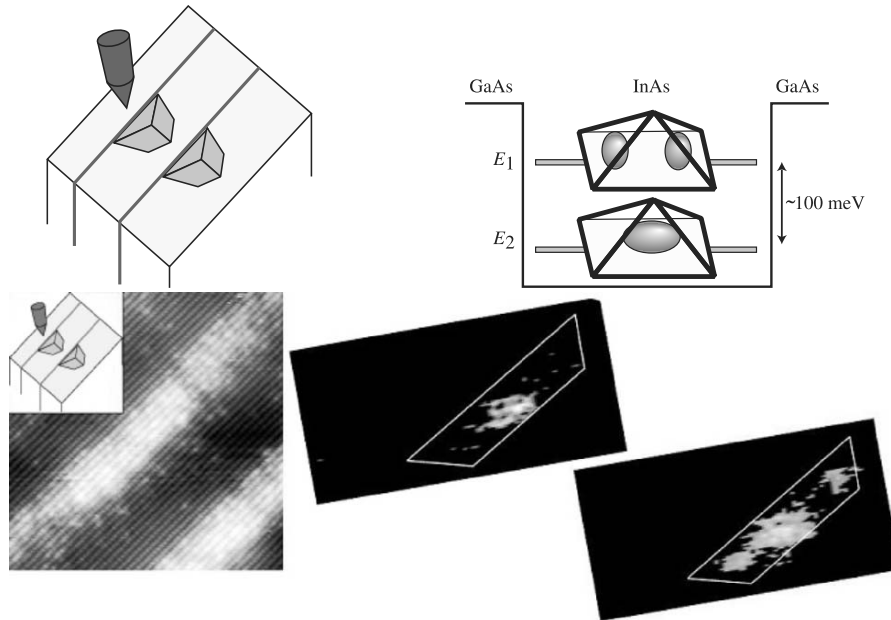
Recall that a QD is a confined system in which the allowed electron energies take discrete values. With these energy levels are associated wave functions with  $s$  symmetry for the ground state level and  $p$  symmetry for the first excited level (see Fig. 3.8d). The squared modulus of the wave function corresponds to the electron density of states. The STM current is proportional to this density



**Fig. 3.7.** Measuring the band gaps of III–V semiconductors. The figure shows the valence bands (negative voltage), the band gap and the conduction bands (positive voltage). For  $V = 0$  V, the Fermi level of the tip is aligned with the Fermi level of the semiconductor. The maximums of the valence band and the edge  $\Gamma$  of the conduction band are indicated by *dashed lines*. The energies corresponding to the band edges are determined by the intersections of the horizontal axis with the tangents to the density of states curve in the region where the latter passes from zero to a nonzero value

of states and it can therefore be measured to determine its spatial symmetry ( $s$  or  $p$ ) by carrying out measurements at different points of the QD. One of the key problems in QD physics today is to determine electron and hole localisation as a function of the QD size and QD–QD coupling.

Figure 3.8a shows a QD (topographic image) and Figs. 3.8b and c are images obtained by carrying out  $I(V)$  measurements at more or less every point of the image and recording the current measured at a given voltage (0.69 V and 0.82 V for Figs. 3.8b and c, respectively). Bright regions correspond to the presence of a current and dark regions to a lack of current. For voltages between 0 and 0.63 V, no current is detected. Above 0.63 V, a current is detected at the center of the image (Fig. 3.8b). For a voltage of 0.82 V, the central feature becomes brighter and new features appear. For voltages above 0.9 V, current is detected at all points of the QD.



**Fig. 3.8.** Spectroscopy on a quantum dot

These observations are explained as follows. For positive voltages (relative to the semiconductor), conduction band states of the semiconductor begin to fill up, as do the quantum states of the QD. For voltages below 0.63 V, there is no current because there are no states between the Fermi level of the semiconductor (located below the bottom of the QD conduction band) and  $eV$  (the Fermi level of the tip). At a bias of 0.69 V, only the ground state level of the island contributes to the tunnel current (Fig. 3.8b, density of states with  $s$  symmetry). At a bias of 0.82 V, the ground state still contributes to the current, but there is also a current due to the first excited state (Fig. 3.8c, additional features with  $p$  symmetry). In the current images, one is therefore observing the squared amplitude of the wave functions for the ground state and first excited state. The difference between these levels (of the order of 115 meV) agrees with theoretical calculation of the electronic structure of a QD with a cleaved surface.

### 3.3.4 Inelastic Tunnel Current

STM imaging and spectroscopy are the basis for far-reaching research on nanostructures, concerning both morphology and electronic structure. However, it is still difficult to identify a molecule adsorbed on a surface if one does not already have a good idea of what is being observed. For this purpose, it is useful to be able to measure the vibration spectrum associated with the



molecule. This type of spectrum includes narrow lines that are characteristic of the vibrational modes of the chemical bonds, and it can be sensitive to the chemical environment of the molecule. Now, using STM, one has direct access to this type of information for a single molecule. Indeed, the inelastic tunnel current is produced when the electrons lose energy as they cross the nanostructure, e.g., molecule, island, localised between the tip and surface. This happens when the energy  $eV$  of the electrons is at least as big as the vibrational energy  $\hbar\omega$  of a chemical bond (of the order of 100–500 meV). A slight increase in the conductance  $dI/dV$  is then observed. To detect it, one must measure the derivative of the conductance, i.e.,  $d^2I/dV^2$ . The absorption lines associated with a chemical bond have widths of a few meV and the resolution of inelastic electron tunneling spectroscopy (IETS) is of the order of  $5.4kT$  [11]. Clearly this type of spectroscopy must be carried out at low temperatures. In general, STM measurements are made in ultrahigh vacuum at temperatures below 10 K.

### 3.4 Tip–Sample Interaction

#### 3.4.1 Manipulation Modes

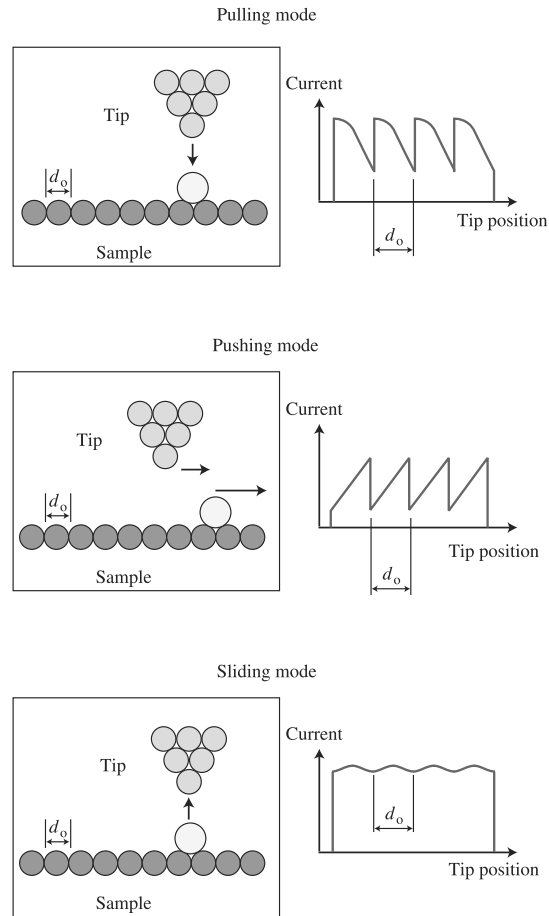
When the tip is brought near the surface, the tip–sample interaction increases and it becomes possible to tear atoms from the surface or manipulate adsorbed atoms (adatoms) by sliding them across the surface. Three manipulation modes are shown in Fig. 3.9 [12].

##### **Pulling Mode**

This mode uses the attractive forces between the tip and the adatom. The tip is positioned above the adatom and then brought towards the surface. The tunnel current increases. The tip is then moved horizontally.  $I$  subsequently falls off until the adatom undertakes a hop towards the tip, while remaining on the surface. The current intensity  $I$  increases once more, and the procedure continues.

##### **Pushing Mode**

This is similar to the last mode, except that it makes use of repulsive forces between the tip and the adatom. The tip is brought towards the surface and moved horizontally towards the adatom. The current intensity  $I$  increases until the tip repels the adatom. The latter jumps to a neighbouring surface site. The current falls abruptly and the procedure continues.

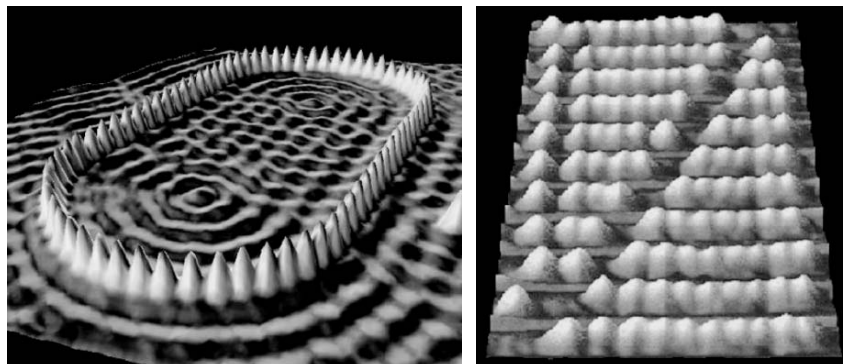


**Fig. 3.9.** Different manipulation modes

### Sliding Mode

In this mode, the forces between the tip and the adatom are attractive, but the tip is so close to the surface that the adatom is attracted onto it. As the tip approaches, the current  $I$  increases and the adatom jumps onto the tip and remains there. When the tip moves parallel to the surface, the current is related to the surface topography as seen by the tip with attached adatom. Finally, the tip is withdrawn and the adatom falls back onto the surface.

These three modes have been studied by Bartels and coworkers using Pb and Cu atoms, and CO molecules, manipulated on a Cu(211) surface. Two illustrations are given in Fig. 3.10. Figure 3.10a shows a quantum corral observed by Eigler, obtained by manipulating iron atoms on a copper surface at 4 K. Apart from the technical feat involved in manipulating atoms one by



**Fig. 3.10.** Manipulations of atoms (*left*) and organic molecules (*right*) using an STM

one and arranging them on the surface, this is a truly magnificent illustration of quantum mechanics. Indeed, the image associated with the surface states represents interference effects within the corral. The focal points of the elliptically shaped corral are clearly visible in the image.

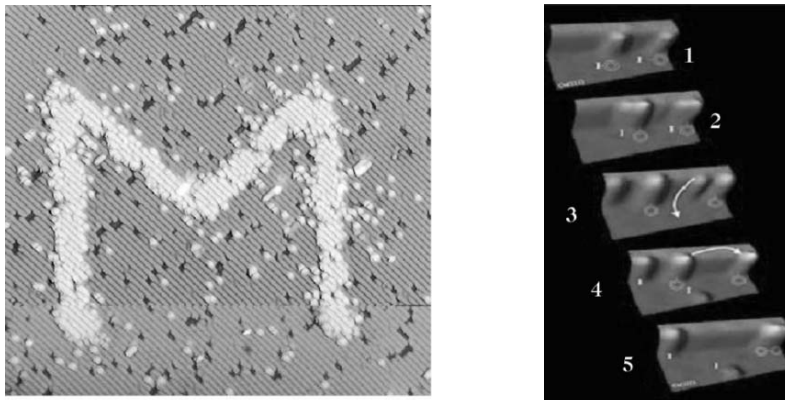
Figure 3.10b shows a molecular abacus in which each counter is an organic nanostructure, in fact, a  $C_{60}$  molecule. This work was achieved by J. Gimzewski on a copper surface. It shows that large molecules can be manipulated, well beyond the size of a single atom.

### 3.4.2 Local Chemistry

We have seen that, when the STM tip approaches the surface, there is a strong interaction which distorts the tunnel barrier. The tip can also interact with the surface at higher currents or voltages. In this case, chemical reactions can occur locally: chemical bonds can be broken by the electric field or the current, due to local heating via inelastic interactions, and chemical reactions can be induced.

#### Local Dehydrogenation

The surface  $Si(100)-2 \times 1$ , which is highly reactive due to dangling silicon bonds, can be passivated by atomic hydrogen in ultrahigh vacuum. Using the STM tip, either with a tip-sample voltage above 4 V or a current of several nA, the surface can be dehydrogenated locally and hence reactivated in a selective manner. Figure 3.11a shows a hydrogenated silicon surface upon which the letter M has been written by local dehydrogenation. The width of the line is 3–4 nm and the area of the image is  $60 \text{ nm} \times 60 \text{ nm}$ . Calculating the number of letters that could be written on  $1 \text{ mm}^2$ , roughly a pin head, one obtains the whole content of the Encyclopedia Universalis (about 400 million characters), as predicted by Feynman in 1959!



**Fig. 3.11.** Examples of local chemistry. *Left:* Selective dehydrogenation of a silicon surface. *Right:* Dissociation of two molecules of benzene iodide (iodobenzene) followed by formation of a diphenyl molecule

### Local Chemical Reactions

The last example illustrates local chemistry combined with manipulation of atoms and molecules (Fig. 3.11b). This work was carried out by a team in Berlin, using a microscope in ultrahigh vacuum and working at 20 K [12]. The manipulation sequence is labelled from 1 to 5:

1. Two iodobenzene molecules appear in the first image with symbols superimposed.
2. After a current pulse, the molecules dissociate.
3. The STM tip moves the iodine atoms away.
4. The two benzene rings are brought together.
5. After another, carefully chosen current pulse, a diphenyl molecule is formed.

This experiment shows the possibility of carrying out local chemistry involving just two molecules.

### 3.5 Conclusion

In this chapter, we have discussed the operating principles and the various uses of the STM, including the imaging mode at atomic resolution and the spectroscopy mode which allows one to determine local electronic structure. Further reading can be found in [13–16]. In these modes, the microscope can observe and measure, but it does not interact with the observed sample. However, for high tip–sample voltages, the tip can interact with the surface to manipulate atoms, extract them from the surface, or even induce local chemistry. In these cases, the STM becomes an active tool for nanofabrication.

However, the STM has its limits, especially with regard to the interpretation of images, which result from a convolution between the topography and the local chemical nature of the sample. For this reason, complementary forms of microscopy have been developed: the atomic force microscope (AFM), sensitive to topography, and the scanning near-field optical microscope (SNOM), sensitive to the interactions of a light wave with the surface. These two microscopes are discussed in Chaps. 4 and 5.

## References

1. Messiah, A.: *Quantum Mechanics*, Dover, New York (2000)
2. Tersoff, J., and Hamann, D.: Phys. Rev. Lett. **50**, 1998 (1983)
3. Bardeen, J.: Phys. Rev. Lett. **6**, 57 (1961)
4. Selloni, A., Carnevali, P., Tosatti, E., and Chen, C.D.: Phys. Rev. B **31**, 2602 (1985)
5. Lannoo, M., and Friedel, P.: *Atomic and Electronic Structure of Surfaces: Theoretical Foundations*, Springer Series in Surface Sciences 16, Springer-Verlag (1991)
6. Gauthier, S., and Joachim, C. (Eds.): *Scanning Probe Microscopy: Beyond the Images*, Les Editions de Physiques (1992)
7. Tersoff, J.: Phys. Rev. B **41**, 1235 (1990)
8. Feenstra, R.: Phys. Rev. B **50**, 4561 (1994)
9. Chen, C.J., and Hamers, R.: J. Vac. Sci. Technol. A **9**, 230 (1993)
10. Feenstra, R.: J. Vac. Sci. Technol. B **7**, 925 (1989)
11. Stripe, B.C., Rezaei, M.A., and Ho, W.: Science **280**, 1732 (1998)
12. Bartels, L., Meyer, G., and Rieder, K.-H.: Phys. Rev. Lett. **79**, 697 (1997)
13. Güntherodt, H.-J., and Wiesendanger, R.: *Scanning Tunneling Microscopy I*, Springer Series in Surface Sciences, Springer-Verlag, Berlin (1992)
14. Bonnell, D.A.: *Scanning Tunneling Microscopy and Spectroscopy*, VCH Publishers (1993)
15. Chen, C.J.: *Introduction to Scanning Tunneling Microscopy*, Oxford University Press, New York (1993)
16. Stroscio, J.A., and Kaiser, W.J.: *Scanning Tunneling Microscopy*, Academic Press, Vol. 27 (1993)

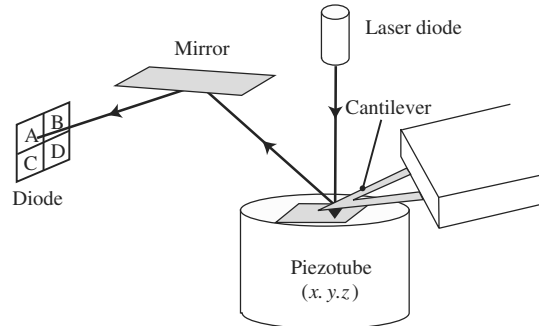
## Atomic Force Microscopy

C. Frétigny

The atomic force microscope (AFM) is undoubtedly the most widely used of the local probe devices. It gives quick access to a wide range of surface properties, including mechanical, electrical, magnetic, and other properties, with good spatial resolution. Furthermore, it can operate in air, vacuum, or solvent. There are certainly many reasons why it can be found in such a large number of research establishments. Not only are the images it provides an invaluable aid in the study of materials, chemistry and physical chemistry, but it is often used for fundamental research, wherein it has contributed to the emergence of nanoscale physics. This type of device also has applications in industry and technology. Due to the relative simplicity of the underlying principles, it is easily integrated into the microelectronic production line, where it fulfills a quality control function. Finally, it constitutes a basic element in promising data storage techniques or the fabrication of miniaturised electronic components. Here too, the AFM has a key role to play in the rise of nanotechnology.

### 4.1 The Device

Figure 4.1 shows schematically how the AFM works. It illustrates a general feature of local probe microscopy, viz., a miniaturised sensor moves near the sample surface. The high degree of spatial localisation in the measured physical quantity is made possible by the small size of the sensor and its close proximity to the surface. The sensor used in AFM is a spring-loaded cantilever, equipped with a tip which interacts with the sample surface. A laser beam reflects off the back of the cantilever, whose deformations under the effects of interaction forces can be measured. The displacement of the spot on a photoelectric cell divided into four dials indicates the deflection and torsion of the cantilever. Displacements are achieved by the deformation of a piezoelectric tube. In Fig. 4.1, the sample moves and the sensor is fixed. In practice, one also finds the opposite system, in which the sensor scans a fixed surface.



**Fig. 4.1.** Schematic diagram of the atomic force microscope. A piezotube displaces the sample located just below the tip carried by a cantilever. Deformations of the bolted cantilever beam are determined by measuring the displacement of the light spot from a reflected laser beam by means of a system of photoelectric diodes. The opposite kind of system also exists, in which the sample is fixed and the cantilever is displaced

The system can work in air, vacuum, or liquid, and it can make measurements at different temperatures.

An image can be formed by recording one or more characteristics of the interacting cantilever beam, e.g., deflection, torsion, amplitude of vibration, etc., at each point of the sample. By means of a servo-system involving the  $z$  displacement of the piezotube, one may also control the distance between the cantilever and the surface in such a way as to hold one of these characteristics at a constant value. The height values then give an image of the sample.

The cantilever and tip are obviously key components of the device. Figure 4.2 shows several images of these components obtained by scanning electron microscope (SEM). As the spatial resolution of measurements is related to the radius of curvature of the tip apex, one seeks to miniaturise the dimensions of the cantilever beam and the tip. Microfabrication processes developed for microelectronics are used to produce them. Consequently, they are usually made from silicon or silicon nitride. Cantilevers with different characteristics are used, depending on the operating mode of the AFM. Reflecting, conducting, or magnetised films are deposited in certain cases. We shall also see that the tip can be chemically modified by tethering or depositing self-assembled layers. Likewise, diamond tips can be used for nanoindentation experiments. In specific applications, silicon beads or carbon nanotubes can be bonded onto bare cantilevers. Table 4.1 summarises typical cantilever characteristics.

## 4.2 The Various Imaging Modes

The main operating modes of the AFM will be described briefly in this section. Later we shall see how to extract, apart from topographical images of the

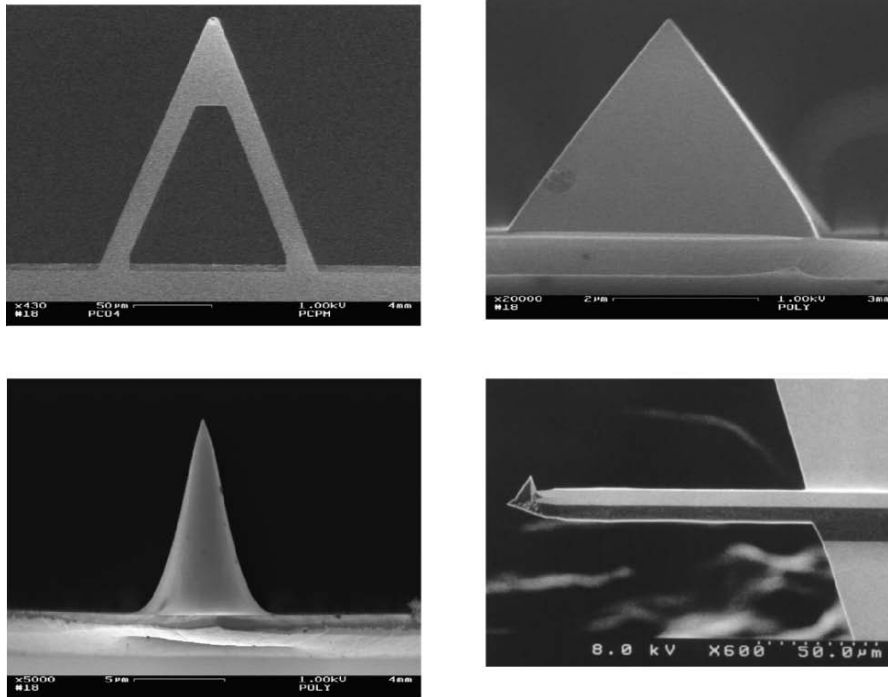


Fig. 4.2. Scanning electron microscope images of different cantilevers and tips

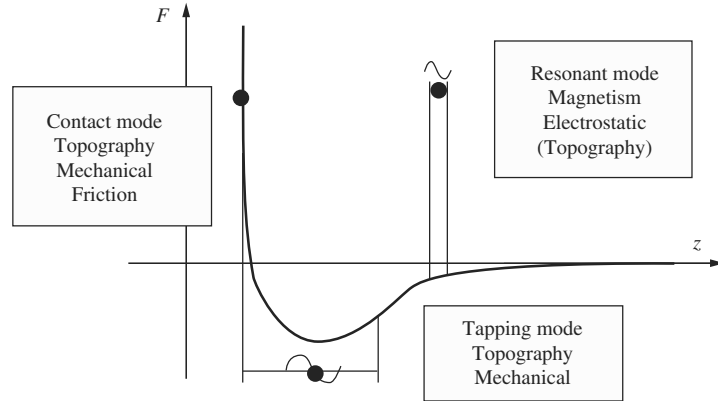
Table 4.1. Typical cantilever characteristics

Shape	Rectangular or V-shaped
Length	50–250 $\mu\text{m}$
Stiffness	0.01–100 N/m
Resonance frequency	10–500 kHz
Radius of curvature of tip apex	2–50 nm
Tip shape	Conical or pyramidal

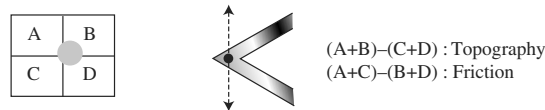
sample surface, supplementary information regarding the physicochemical or physical properties of the surface.

Suppose to begin with that the sample is made from a non-deformable material. The tip-sample interaction is represented in Fig. 4.3 by a Lennard-Jones-type interaction force, in which the interaction is attractive at large distances (typically, beyond a few tenths of a nanometer) due to van der Waals forces, and repulsive at very short range due to the impenetrability of the electron clouds associated with the two surfaces.





**Fig. 4.3.** Three AFM operating modes are located by *black dots* in a diagram showing the interaction force between tip and sample. In the tapping and resonant modes, the cantilever vibrates close to its resonance frequency, whereas the contact mode is quasi-static



**Fig. 4.4.** To measure friction forces, the sample is scanned perpendicularly to its longitudinal axis. Torsion induced by these forces is revealed by a horizontal shift of the laser spot on the photodiodes

### Contact Mode

Historically the first form of AFM, this mode operates close to the repulsive edge of the potential. In this sense, the tip can be said to actually touch the surface. However, with certain samples, wear and tear, and deformations caused by the tip impair image quality. This mode is quick and easy to use, and it is often combined with simultaneous measurements of friction, adhesion or contact stiffness, described below.

### Friction Mode

This is friction force microscopy (FFM) or lateral force microscopy (LFM). In contact mode, the sample can be scanned perpendicularly to the cantilever axis. In this case, friction forces cause torsion at the end of the cantilever beam, thus displacing the laser spot in the horizontal plane (see Fig. 4.4). By reading this displacement, one obtains a measure of the friction forces which can be a true measure if the system has been suitably calibrated.

### Resonant Mode

In this operating mode, which could be described as the linear resonant mode, the cantilever is made to oscillate at its resonance frequency, ‘far’ from the surface and with ‘small’ amplitude. The terms ‘far’ and ‘small’ are of course relative and will be specified more precisely in the section dealing with resonant modes. The gradient of the interaction force shifts the resonance frequency of the cantilever. As the tip oscillates relatively far from the surface, a certain degree of spatial resolution is lost, so that this mode is not generally used for topographical studies. However, it can serve to analyse long-range electric or magnetic forces, by using conducting or magnetic tips, respectively.

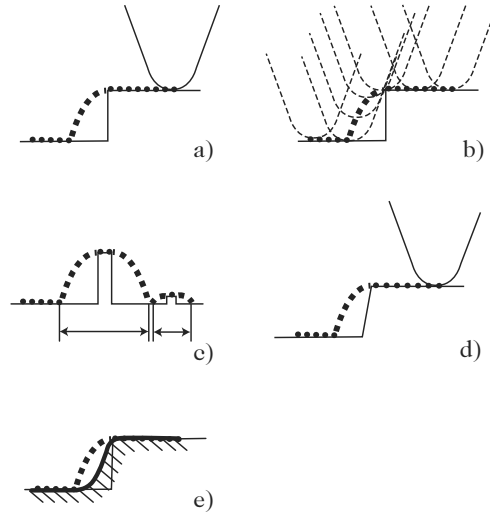
### Tapping Mode

This mode, also known as intermittent contact mode, is a nonlinear resonant mode in which oscillation amplitudes are larger and the mean position of the tip is closer to the surface. In each cycle, the tip can be said to brush against the repulsive wall of the surface. It is more difficult to analyse this operating mode, which is widely used to determine sample topography. Forces applied to sample surfaces can be extremely small and contact times so short that almost no friction force occurs. One can therefore avoid deformation of the sample and the kind of wear that is always possible in contact mode. Moreover, due to the brevity of contact, there is no time for adhesive effects to arise. The size of the contact region is very small, even on highly deformable samples, and this leads to good lateral resolution. When the sample height is servo-controlled at a constant amplitude, the phase difference between the excitation and the oscillation of the cantilever beam characterises dissipation from the system. Phase images can thus reveal slight heterogeneities in the sample surface, corresponding to different viscoelastic, adhesive or wetting properties.

## 4.3 Image Resolution

Very early on, images obtained by contact mode AFM were able to show the crystallographic periodicity of certain surfaces, and this contributed significantly to the success of the method. In this case, the mechanism underlying contrast formation, probably caused by the jerky rubbing motion of the tip, would only appear to be possible on rather special kinds of sample with a certain degree of surface roughness on the atomic scale. It is precisely the periodic arrangement of the surface that leads to the formation of the image, so that one could not pick out a one-off defect, for example.

Recently, dynamic mode resonant techniques have made it possible to visualise surface atoms under ultrahigh vacuum conditions. This operating mode, which yields high quality data, comparable with those obtained by scanning tunneling microscopy (STM), is still poorly understood and is currently under

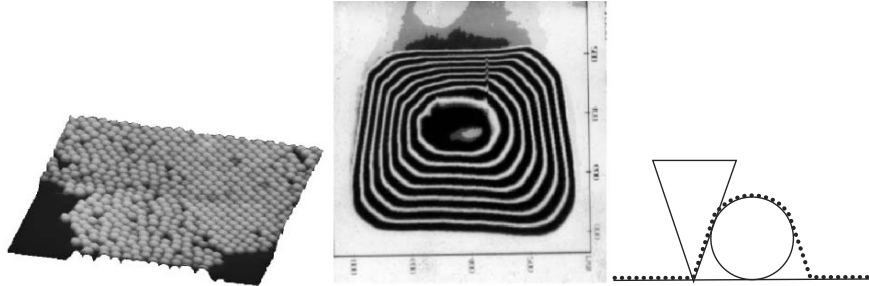


**Fig. 4.5.** Finite size effects due to the tip. The measured trajectory of the tip is shown by *black squares*. Successive positions of the tip as it passes over the step (a) are shown in (b). Two objects of different heights lead to different lateral extensions as shown in (c). A slightly sloping step (d) cannot be distinguished from the step in (a). Finally, a data processing technique based on analysis of the tip positions in (b) optimises the image of the step that can be obtained with this tip (e)

active investigation. However, AFM is generally used to image surfaces on a mesoscopic scale. It is the contact mechanics that determines the resolution for highly deformable materials, e.g., with Young's modulus below 100 Mpa. In the case of only slightly deformable samples, the vertical resolution of images is generally very good. It is only limited by the sensitivity with which the amplitude or deflection are detected (of the order of 0.1 nm) and the accuracy with which the vertical displacement of the piezoceramic is controlled (of the same order of magnitude). For topographical studies, one can say that the vertical resolution with this method is better than the interatomic distance, whereas the lateral resolution has to be treated with great caution.

To simplify, we shall suppose here that the AFM operates in contact mode as a perfect tactile sensor passing over a non-deformable surface. However, similar conclusions can be drawn in other modes. What is the resolution of this imaging mode? We shall see that the answer is not as simple as in optical or electron microscopy, for example.

The diagram in Fig. 4.5a shows the path followed by an AFM tip on one scan over a vertical step. The broadening and distortion of the shape of the object are due to the bulkiness of the tip (Fig. 4.5b), which feels the presence of the step before its apex reaches the position vertically above the step edge. The point of contact between tip and surface remains at the step edge until the tip apex has passed through the vertical above this point, with the image



**Fig. 4.6.** *Left:* An island composed of a monolayer of monodispersed silicon beads. *Centre:* Single bead imaged with a pyramidal tip. The distortion caused by the tip geometry is clearly visible. *Right:* Schematic of image formation

arising from the flanks of the tip. Beyond this, the trajectory is once again dictated by a contact between the tip apex and the surface. The picture here is two-dimensional. In three dimensions, it is clear that other situations can arise with regard to the tip-sample contact point, according to exactly the same principle.

The geometric effects due to the finite size of the tip complicate any discussion of resolution. From Fig. 4.5c, it is clear that the lateral broadening of an object of given width will depend on its height. The lateral resolution of AFM cannot be described by an instrumental profile as it can in optical microscopy, for example. In fact, the imaging process is not linear. Figure 4.5d shows that a slightly sloping step will give exactly the same image as the vertical step in Fig. 4.5a. This means that information can be completely lost by the imaging mechanism, in a way that would not happen with a convolution. Although the term is not strictly applicable, one still speaks of the tip convolution.

Figure 4.5e uses the successive positions of the tip from Fig. 4.5b to determine an optimal boundary beyond which the actual step surface cannot be located. This data processing technique can be used to refine the resolution of images acquired by a tip of known shape. In order to discover this shape, one carries out the opposite investigation on a rough sample surface: at each point of the image, no part of the tip can be located in the half-space below the recorded surface. Hence, by successive elimination of known regions, one can reconstruct the shape of the tip. Several algorithms have been put forward to achieve these two aims.

Figure 4.6 illustrates the broadening effect produced by a pyramidal tip. The first image ( $5\ \mu\text{m} \times 5\ \mu\text{m}$ ) shows a monolayer island of silicon beads. It is known by other means that each bead is perfectly spherical. The magnification of a single bead shown by contours in the second image ( $900\ \text{nm} \times 900\ \text{nm}$ ) reveals a distinctly pyramidal shape. The diagram on the right shows the mechanism leading to broadening in this case.

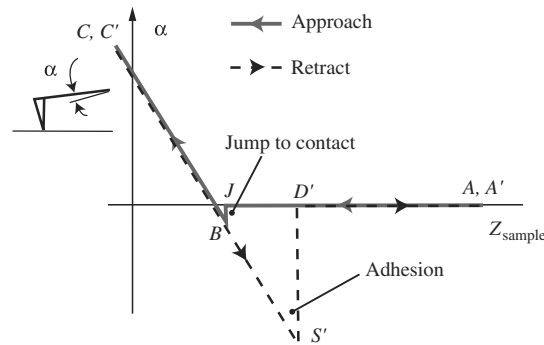
The resolution cannot be defined by a simple number in contact AFM. The finite-size effect due to the bulkiness of the tip increases as the tip gets

blunter and the aspect ratio decreases. For example, a broken tip will reveal all the smaller details on the sample surface by a single characteristic shape. A broken tip can often be identified through this behaviour.

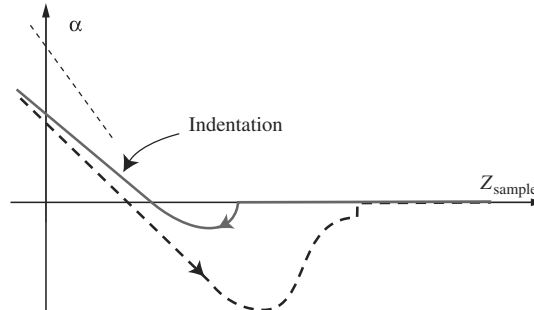
#### 4.4 Contact Mode: Topography, Elasticity and Adhesion Imaging

The contact mode can be described on the basis of the so-called force curve, which represents the variation of the cantilever deflection as a function of the sample height (tip-sample separation), as exemplified in Fig. 4.7. Without vibrating the cantilever, the vertical position of the sample is varied and the cantilever deflections are recorded. The approach paths, moving towards contact (from right to left in the figure), differ from the retract paths, in which contact is broken (from left to right). The graph can be analysed into several parts:

- **Approach.** Far from the sample surface, the interaction forces are very weak and there is almost no deflection of the cantilever. This is the horizontal part of the curve on the right (return trip between  $A$  and  $J$ ). In vacuum, air, and sometimes even liquids, the non-contact tip-sample interaction is attractive and causes a slight downward deflection of the cantilever (negative  $\alpha$ ) which is generally barely visible. During approach, this slightly deflected position becomes unstable at  $J$  and the tip jumps to contact at  $B$ . The corresponding instability is revealed by the vertical



**Fig. 4.7.** Force curve on an ideal non-deformable material. The deflection of the cantilever beam is graphed as a function of the vertical position of the sample during an approach-retract cycle. Once contact is established at  $B$ , the deflection increases in proportion with the rise of the sample surface, where  $BC$  corresponds to approach and  $C'S'$  to retraction. As the tip moves away, contact is only broken when the adhesive forces can no longer withstand the separating force exerted by the cantilever (point  $S'S$ )



**Fig. 4.8.** Force curve for a deformable material. The *dashed straight line segment* shows the slope of the linear contact region which would be measured on a perfectly rigid material like the one represented in Fig. 4.7

jump  $JB$  of the curve during approach. If the sample is further raised toward the tip, for a very rigid material, the deflection will increase linearly with the sample height ( $BC$ ).

- **Retraction.** The force curve begins by retracing the approach path. However, it goes beyond the zero force position and even the point where the jump to contact occurred. This is due to adhesion and is indicated by the curve  $C'S'$ . One must in fact exert a separating force on the contact to break it. Until the breaking point is reached, the trajectory continues along the straight line characterising contact. When the breaking point  $S'$  is reached, the cantilever moves back to the very slightly deflected position at  $D'$ .

Adhesion is thus manifested through hysteresis in the force curve. It is caused by several factors: van der Waals forces, as one would expect, but also electrostatic forces and capillary forces in liquids, etc. These interactions are then affected by the pH, ionic forces, and so on. On the basis of these comments, it is easy to see that AFM is highly sensitive to the physicochemical properties of surfaces.

Operation of this instrument can also be affected by the mechanical properties of the sample. Figure 4.8 shows deformations one might expect from a rather deformable material. Once contact has been established, the tip is pressed against the material by the elasticity of the cantilever. It may then penetrate into the material, so that the recorded deflection will be less than would be obtained on a perfectly rigid sample, as indicated by the dashed straight line segment in the figure. As the sample is raised, the increase in the deflection is thus slower and is characteristic of the stiffness of the contact. (A simple model can be made by adding the stiffness of the cantilever beam and the stiffness of the contact in series.) Likewise, the contact may not break abruptly, since the material may exhibit some degree of creep before complete rupture occurs. Moreover, if the sample is viscoelastic, the curve will distort

when one modifies the operating frequency, tip shape, maximal penetration in the material, and so on.

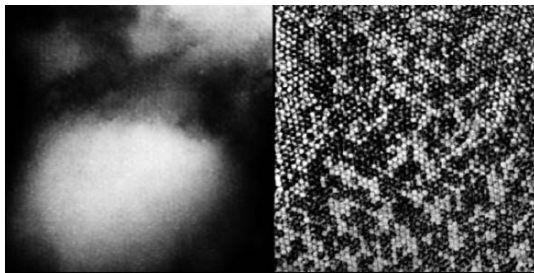
Consequently, if the sample height is periodically modulated, the deflection response to the modulation frequency is a measure of the contact stiffness. This is called elasticity mode. Likewise, a periodic modulation can be imposed in such a way that the tip goes through the minimum characterising adhesion. The characteristics of the deflection signal obtained in this way then constitute a signature for adhesion. One can also record the whole force curve at each point of the image and then process the data, but the huge amount of information acquired in this way makes this approach too time-consuming. It is preferable to retain only the main characteristics of the curves through appropriate signal processing. This is what is done in the pulsed force mode.

An important characteristic of all these operating modes is that they can often be used at the same point of the sample, without changing the cantilever beam and sometimes even simultaneously. For example, an elasticity or friction image combined with a height image often reveals sufficient detail to understand the physicochemical structure of a sample, without necessarily needing to be concerned with the dissipation or tribological properties in their own right. Shapes and sizes brought out by the contrast are used directly. It is not then necessary to calibrate the instrument or to provide a precise model of operation. To avoid certain experimental artifacts or obtain more quantitative measurements, on the other hand, the operation of these different modes must be analysed in more detail.

#### 4.4.1 Friction Mode

We have seen that the lateral force is measured directly from the cantilever torsion, which is in turn detected as a displacement of the laser spot in the horizontal plane.

Figure 4.9 shows a direct application of friction mode imaging. A sample of latex beads has been heated above the glass transition temperature of the



**Fig. 4.9.** AFM images of the surface of a latex film. *Left:* Surface topography. *Right:* Simultaneously acquired image revealing local friction variations associated with physicochemical properties of the material

polymer in order to create an apparently homogeneous material by interpenetration of molecular chains. The height image shows nothing more than a certain roughness on the micrometer scale, as would be found for many polymers. However, the simultaneously acquired friction image shows a highly regular arrangement of domains measuring some 100 nm across, the size of the latex beads. The material thus has a memory of its preheated structure, and we have the demonstration that homogenisation has not been perfect. Another point is worth noting: two types of bead can be made out in the friction image. Their tribological behaviour differs and their physicochemical properties too (viscoelastic modulus, surface energy, or other). Even though one cannot proceed much beyond the simple observation without further experiments, the friction mode has at least brought out an important feature of the material in a simple and direct way.

## 4.5 Resonant Modes

The cantilever, effectively a beam clamped at one end, is a good mechanical oscillator with a low level of dissipation, mainly caused by the viscosity of the ambient medium. Quality factors of around 400 are common in air, several tens of thousands in ultrahigh vacuum, and 10 in water. The resonance frequency is modified by tip–surface interactions. This idea is applied in the AFM resonant modes. Two techniques are used:

- One can drive the cantilever at a fixed frequency close to its resonance frequency and follow variations in the amplitude, and possibly also the phase. This is the method most commonly used for experiments carried out in air. In this case, the sample height is servo-controlled at a constant amplitude. This technique is sometimes called amplitude modulated AFM (AM-AFM).
- One can also set up a phase-locked loop which holds the vibration amplitude and phase difference at pre-assigned values. The servo-controlled system then begins to self-oscillate and its resonance frequency is continuously monitored. This is dynamic mode or frequency modulated AFM (FM-AFM).

We shall not discuss the second method in further detail here. It is mainly used in ultrahigh vacuum experiments where the natural quality factor is very high, so that the oscillator has a very long reaction time (of the order of  $Q/\omega_0$ ). However, the discussion below applies more or less directly to this scenario. The reference [1] provides a recent and very complete review of oscillating methods in force microscopy.

### 4.5.1 General Principles

In order to describe the operation of resonant modes, the general problem that has to be solved concerns the vibrational modes of a beam that is clamped at



one end, whilst the other is subjected to a force field. Although a complete study is possible, we shall not present one here. The high quality coefficients observed experimentally allow one to restrict the analysis to a single mode for which the equation of motion reduces, to a very good approximation, to that of a harmonic oscillator subjected to a force field. The appropriate equation for this system is

$$\ddot{x} + 2\beta\dot{x} + \omega_0^2 x = \gamma \cos \omega t + \frac{f(D, t)}{m}, \quad (4.1)$$

where  $x$  is the position coordinate of the oscillator, which is in the present situation the displacement of the tip from its equilibrium position,  $\omega_0$  is the resonance frequency of the oscillator,  $\gamma$  is the amplitude of the excitation at frequency  $\omega$ ,  $\beta$  is a dissipation term such that the quality factor is given by  $Q = \omega_0/2\beta$ , and  $m$  is the effective mass of the oscillator, determined by  $\omega_0 = k/m$ , where  $k$  is the cantilever stiffness. The function  $f(D, t)$  is the tip-sample interaction, where  $D$  is the tip-sample separation when the cantilever is not deflected.

The simplest case occurs when the interaction force depends only on the tip-sample separation  $D+x$ . More complex behaviour is observed if dissipative behaviour comes into play due to adhesion, viscoelasticity, or capillarity.

Even in the very simple case described by  $f(D, t) \equiv F(D+x)$ , the equation governing the system is not generally linear:

$$\ddot{x} + 2\beta\dot{x} + \omega_0^2 x = \gamma \cos \omega t + \frac{F(D+x)}{m}. \quad (4.2)$$

#### 4.5.2 Linear Resonant Mode

It is easy to describe the linear resonant mode, which corresponds to a non-dissipative interaction and a very low amplitude oscillation far from the sample surface ( $x \ll D$ ). Expanding the interaction to first order in  $x$ , (4.2) gives

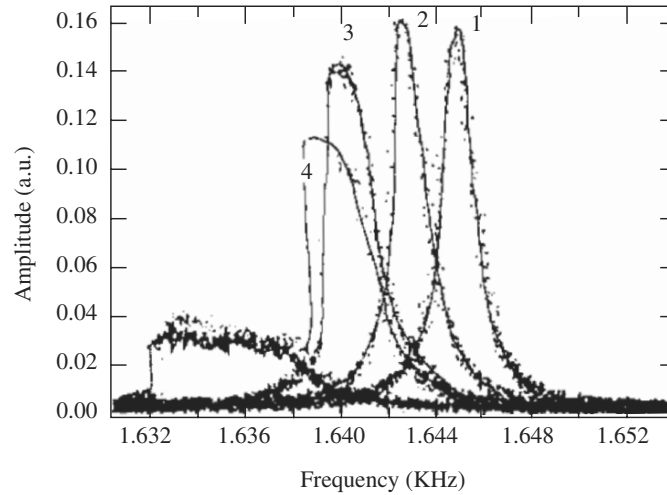
$$\ddot{x} + 2\beta\dot{x} + \omega_0^2 x = \gamma \cos \omega t + \frac{F(D)}{m} + \frac{F'(D)}{m} x,$$

where  $F'(D)$  is the gradient of the force at the central position of the oscillation. The constant term shifts the rest position of the tip  $F(D)/k$ , which is generally negligible compared with the oscillation amplitude. We thus obtain the new equation, expressed relative to the new mean position,

$$\ddot{x} + 2\beta\dot{x} + \left[ \omega_0^2 - \frac{F'(D)}{m} \right] x = \gamma \cos \omega t.$$

This is the equation of a harmonic oscillator whose resonance frequency  $\omega'_0$  satisfies

$$\omega_0'^2 = \omega_0^2 \left[ 1 - \frac{F'(D)}{k} \right].$$



**Fig. 4.10.** Resonance spectra of a cantilever at various distances from an MgO surface. Curves numbered from 1 to 5 correspond to tip-sample separations of 80, 60, 50, 40 and 10 nm, respectively. Note the asymmetry of the peak at shorter distances [2]

The resonance frequency of the cantilever is thus shifted by the gradient of the interaction. As the magnitude of the interfacial forces decreases with distance, an attractive (repulsive) interaction causes a reduction (increase) in the resonance frequency of the system, as one would expect qualitatively.

This linear resonant method has been used to carry out measurements of long-range interfacial forces. It is now commonly used to obtain electrostatic or magnetic images of surfaces. Several applications will be described in Sect. 4.6.

The above analysis assumes that a first order expansion of the interaction is adequate. This interpretation is borne out by recordings of the resonance spectrum, which is a characteristic feature of a harmonic oscillator. In practice, the tip must oscillate rather far away from the sample surface for this linear approximation to be valid. At shorter distances, the resonance peak is distorted and a more complete analysis of the equation is in order (see below). Figure 4.10 shows the resonance spectra of a tungsten tip close to an MgO surface [2]. The tip-sample separations are 80, 60, 50, 40 and 10 nm in spectra 1–5. One first observes a simple shift of the peak, but then at smaller separations the resonance spectrum becomes asymmetrical. This behaviour, characteristic of nonlinear oscillators, is discussed in the next section.

#### 4.5.3 Nonlinear Resonant (Tapping) Mode

If the tip vibrates close to the surface, or if the amplitude of vibration is large, a first order description of the interaction is no longer adequate: the oscillator

is nonlinear. In practice, however, the deflection remains quasi-sinusoidal. In fact, due to the high quality factor of the resonance, nonlinearities do not lead to anharmonicity in the response. Instead, they manifest themselves in the way the response depends on the control and interaction parameters. Returning to (4.1), viz.,

$$\ddot{x} + 2\beta\dot{x} + \omega_0^2 x = \gamma \cos \omega t + \frac{f(D, t)}{m},$$

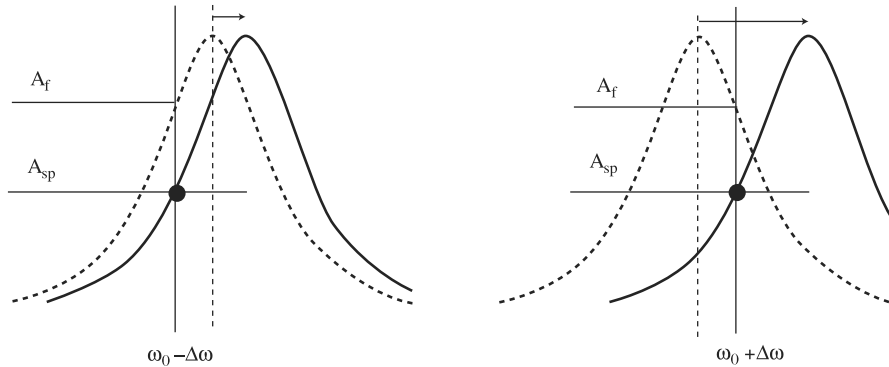
we seek a solution of the form  $x = A \cos(\omega t - \varphi)$ . In this case, the interaction is a periodic function of time that can be represented by its Fourier series expansion. A complete analysis of this operating mode would go beyond the scope of this discussion. However, one important consequence of this description concerns the analysis of dissipation due to the tip-sample interaction. The energy dissipated by the interaction over one cycle, denoted by  $U_{\text{diss}} = \oint f(D, t) dx$ , can be expressed as

$$U_{\text{diss}} = Am\gamma \left( \frac{2A\beta\omega}{\gamma} - \sin \varphi \right). \quad (4.3)$$

Hence, if the amplitude is held constant when an image is scanned, the phase offset  $\varphi$  between the excitation and the oscillation of the cantilever constitutes a measure of the local dissipation  $U_{\text{diss}}$  due to the tip-sample interaction, all other parameters in (4.3) remaining constant.

The experimenter can draw several useful conclusions from this analysis of the tapping mode. The nonlinearity of the system makes it highly sensitive to local variations in the physicochemical properties of the sample surface. One suggestion has been to use the slope of approach-retract curves (analogues of the force curves for this mode) to estimate the elastic modulus of the material beneath the tip. One may discuss the local properties of the material on the basis of a point-by-point study of the image. However, once again, the converse problem is difficult to solve: the various models tend to provide only qualitative answers to questions concerning the underlying causes of image contrast.

Although the tapping mode is a nonlinear operating mode, the linear behaviour of the shift in resonance frequency can be very useful for a rough analysis of what is happening. Hence, if we seek to understand the effect of the choice of working frequency on the tip-sample interaction, we may refer to this model. Figure 4.11 compares experimental situations for the same servo-controlled amplitude (the set point)  $A_{\text{sp}}$  and working frequencies  $\omega_0 \pm \Delta\omega$  symmetrically placed with respect to the resonance frequency of the free oscillator, leading to the same free amplitude  $A_f$ . For a rather repulsive interaction, the resonance peak is shifted towards higher frequencies. An adjustment of the free resonance leftwards to  $\omega_0 - \Delta\omega$  will satisfy the condition for servo-controlling the amplitude for a smaller shift than the one that would be required for an adjustment to the right. The corresponding interaction will thus be reduced,



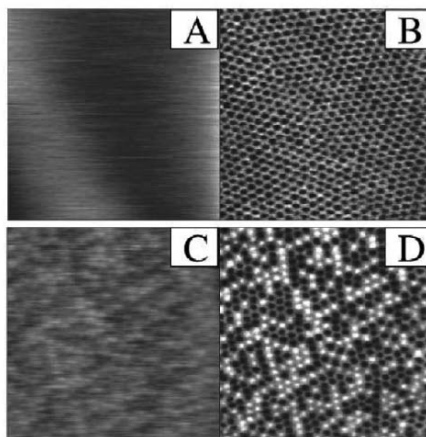
**Fig. 4.11.** Two symmetrical adjustments on either side of the resonance frequency lead to very different operating interactions. For the same working amplitude, an excitation to the right of the resonance peak requires a bigger interaction than an excitation of the same amplitude to the left of the peak

and one can image in a ‘softer’ repulsive mode. This should help to reduce tip damage.

Simplifying somewhat, the free vibration amplitude of the cantilever beam constitutes a measure of the energy contained in the oscillator. The smaller this amplitude is, the more the operation of the system is affected by non-contact interactions or adhesion effects. On the other hand, large amplitudes will be more sensitive to local mechanical differences at the sample surface. These deliberately simplified considerations have to be adapted to suit the case. For example, the radius of the tip apex is an important factor which fixes both the strength of non-contact forces and the contact stiffness.

Given the high sensitivity of the oscillator with regard to the details of the tip-sample interaction, phase images tend to incorporate a great deal of information. Of the two parameters which describe the oscillation (amplitude and phase), the amplitude is held constant by the feedback loop and tends to characterise the topography, whilst the phase evolves in a way that depends on the physical parameters of the interaction, such as viscoelasticity, presence of contamination, adhesion, and so on. The precise cause of the phase contrasts generally remains unknown, but the information gathered can usefully contribute to one’s understanding of a sample surface. Figure 4.12 shows height and phase images of a block copolymer sample. Although the topography is rather flat, differences between local mechanical properties are clearly revealed in the phase image.

Another consequence of this high sensitivity to the details of the interaction is the possibility of artifacts in height measurements. For materials with different viscoelastic moduli or adhesion properties, the change in the form of the interactions does not guarantee that the feedback device will exactly correct for the height differences. Hence, height measurements on DNA molecules



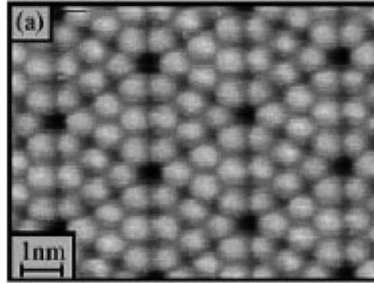
**Fig. 4.12.** Topography (A) and phase offset (B) measured simultaneously on a sample of poly(butadiene-*b*-ethyleneoxide) ( $500 \times 500 \text{nm}^2$ ). B and C show the same quantities after partial crystallisation. 12-nm spheres crystallise independently into a hexagonal structure. *Dark regions* correspond to regions of melt polymer in which viscoelastic dissipation is high [14]

deposited on mica vary significantly depending on the operating conditions chosen for the observation (set point amplitude, drive amplitude, working frequency, etc.): the major differences in chemical nature between the substrate and adsorbate give rise to very different oscillator behaviour, and these in turn lead to the same oscillation amplitude for very different tip altitudes. In practice, these effects are relevant when measuring very small heights and when dealing with highly dissimilar materials. In other situations, they are negligible. Such measurements can be confirmed by analysing the contrast as a function of the operating point (or approach–retract curves, which amounts to the same thing).

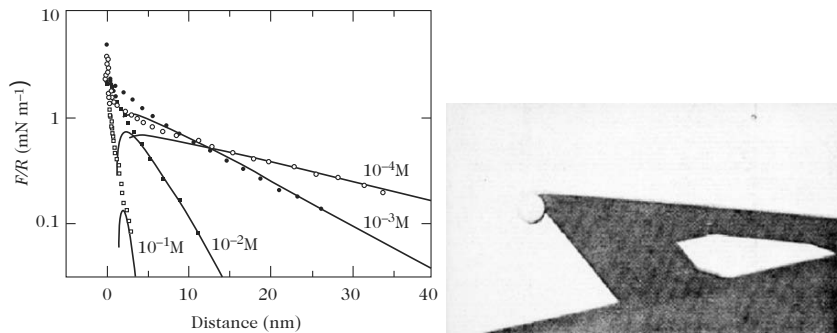
### Tapping Mode in Liquid and in UHV

In a liquid, resonance conditions of the cantilever are modified by the surrounding medium. The added mass, i.e., the mass of liquid that the beam must displace in order to vibrate, reduces the resonance frequency, e.g., by a factor of about 2 in water, while the viscosity diminishes the quality factor in proportion to the viscosity ratio, e.g., by a factor of about 50 in water. In order to work in such conditions, several excitation methods have been suggested, including vibration of the cell for a liquid, and mechanical or magnetic excitation. These techniques have found many applications to problems in biology.

In ultrahigh vacuum (UHV), resonance quality factors are very high ( $\sim 10^5$ ). Characteristic times for the oscillator to reach equilibrium would



**Fig. 4.13.**  $10 \times 7 \text{ nm}^2$  image of the surface of a silicon (111) sample with  $7 \times 7$  reconstruction, as obtained by dynamic mode AFM in ultrahigh vacuum [4]



**Fig. 4.14.** Interaction force vs. distance for a silicon sphere and a surface of the same material, as obtained by analysing force curves. The bead was bonded to the cantilever at the position of the tip (*right*) [5]

prevent excitation at a fixed frequency. For this reason, and apparently also for reasons of sensitivity, most work has been done using a phase-locked loop. In fact, it is often by using this method that genuine atomic resolution has been obtained. An example is shown in Fig. 4.13, which is an image of the  $7 \times 7$  reconstruction of a silicon (111) surface.

## 4.6 Force Measurements

The ability of AFM to make local measurements of weak forces, either contact or non-contact, quickly drew the attention of a wide cross-section of the scientific community: from cellular biologists concerned with questions of mechanical stimulation, to physicochemists faced with fundamental problems of interfacial interaction potentials. The first approach was to use force curves for these explorations, as discussed above. To illustrate this application and also the use of the linear resonant mode, several examples will be selected from the literature.

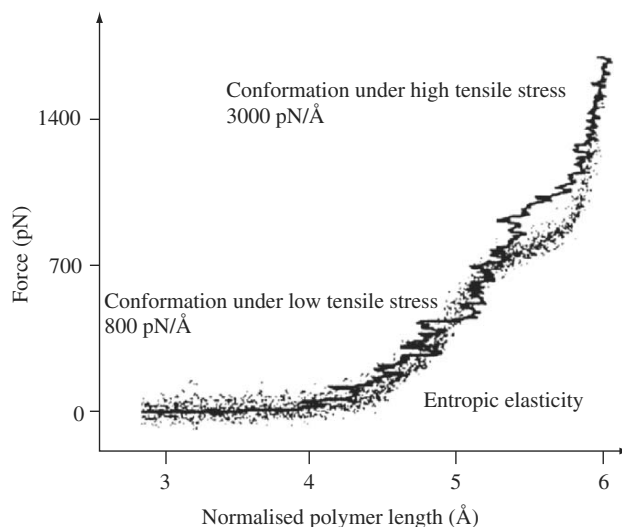
#### 4.6.1 Non-Contact Measurements

Force curves can be used to find a formula for the forces between the tip apex and the sample surface. In a liquid, electrostatic interactions arising from the presence of ions are revealed by analysing such curves. The DLVO model, due to Derjaguin, Landau, Verwey and Overbeek, which purports to describe this effect, has been confirmed for a great many combinations of tip and sample. Figure 4.14 shows the interaction between a silicon bead of diameter  $3.5\ \mu\text{m}$  and a surface of the same material, for various concentrations of NaCl. Observations have been fitted to the DLVO model (continuous curves).

#### 4.6.2 Elasticity and Adhesion Measurements on a Single Molecule

Having treated the tip, specific molecules tethered or adsorbed onto the sample surface can be selectively adsorbed by a kind of fishing technique. The point on the force curve where the adhesion force gives way corresponds to accidents relating to typical features of the elongation and the rupture of adsorption of individual molecules. Figure 4.15 shows the results of an experiment carried out with dextran polymers. Specific bonds can form between molecules attached to the tip and functional groups in the dextrans.

Since the end-to-end length (or contour length) of the polymer is arbitrary, the curves have been renormalised to a fixed length. The experimentally



**Fig. 4.15.** A chemically modified tip can selectively adsorb molecules attached to the sample surface. With dextran polymers, the force–distance curves thus obtained reflect the elasticity of the molecules. As the end-to-end length of the polymers is arbitrary, the curves must be renormalised to a single length to reveal the mechanisms governing molecular elasticity [6]

determined relationship between the force and the elongation is shown in Fig. 4.15. In the same figure, a numerical simulation leads to an interpretation of the two observed regimes in terms of entropic and enthalpic elasticity. The fact that the measurements can be repeated on the same molecule confirms the elastic character of the mechanism. These experiments require a highly stable instrument and extremely good signal-to-noise ratio in order to measure forces of the order of 100 pN in a repeatable manner.

## 4.7 Magnetic and Electrical Measurements

As we saw above, long-range forces can be investigated in linear resonant mode. One generally uses a double scanning technique, where the first scan in intermittent contact determines the surface topography, and a second scan carries the tip over the same points but at a constant distance above the surface, typically 10–20 nm, which eliminates van der Waals forces. In some cases, the cantilever deflection can simply be recorded in static mode.

### 4.7.1 Magnetic Measurements

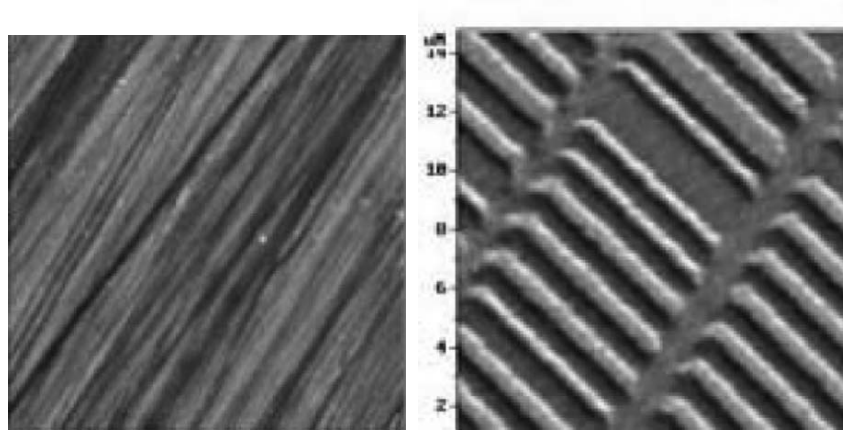
Magnetic tips can be magnetised either along or perpendicularly to their axis using a magnet. Magnetic AFM is used to analyse domain structures of alloys, small particles, or magnetic contacts, not to mention vortex lattices in superconductors. It is also commonly used to study recording systems, such as magnetic tapes, hard or magneto-optical disks, recording heads, and so on (see Fig. 4.16). The energy of interaction between tip and sample can be viewed as the energy of the field produced by the sample on the dipoles of the tip, although this is of course the same as the energy of the field produced by the tip on the sample. There is therefore a mutual interaction which can perturb measurements when magnetic rigidities are low in either the tip or the sample. We should also mention attempts to obtain mechanical detection of nuclear magnetic resonance (NMR) for a small number of spins.

### 4.7.2 Electrical Measurements

Electrical measurements are important in microelectronics. However, like the other forms of local characterisation, they can prove invaluable for revealing physicochemical differences in sample surfaces. They can also help with difficult problems such as triboelectricity, when geometric conditions allow it, i.e., no roughness on the scale of the tip–sample contact.

Several types of electrical measurement are possible with AFM. For resonant modes, one uses a conducting tip which interacts with the sample via Coulomb forces. Doping of semiconductors, presence of localised charges, electrical polarisation, or work functions for removing electrons from the surface



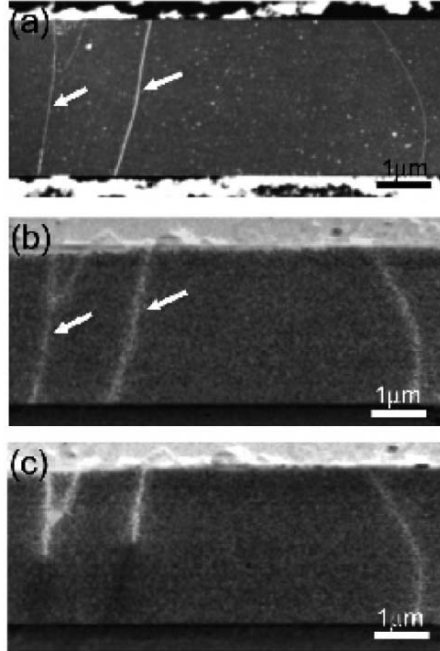


**Fig. 4.16.** Simultaneously acquired  $14 \times 14 \mu\text{m}^2$  images of a hard disk. *Left:* Topography. *Right:* Magnetic image obtained by measuring the phase shift of the oscillation when the tip is at a mean height of 50 nm above the surface. See [www.ntmdt.ru/applicationnotes/MFM](http://www.ntmdt.ru/applicationnotes/MFM)

can all be revealed by a change in the resonance characteristics of the cantilever. One can directly measure the gradient of the force on the tip when it is held at a reference potential, as we saw in Sect. 4.5.2, or when it is excited by an alternating voltage.

In the first case, electric force microscopy (EFM) is particularly useful for studying components, such as working transistors. The passivation layer does not prevent the electric field produced by polarised parts from interacting with the tip. This brand of AFM is also widely exploited in microelectronics, where it serves to analyse defects such as cut tracks, charge accumulation regions, and so on. Figure 4.17 illustrates the method for carbon nanotubes electrically connected to two electrodes.

When a potential difference is applied between the tip and sample, it is the electrical characteristics of the interaction that create the resonance. This is known as Kelvin force microscopy (KFM). Since the applied force is capacitive, it depends quadratically on the potential. For a potential difference of the form  $V + v \sin \omega t$ , the force will include, apart from its mean value, a contribution at the excitation frequency  $\omega$ , with strength proportional to the product  $vV$ , and another at twice the excitation frequency with strength proportional to  $v^2$ . For example, the cantilever can be driven at a frequency equal to half its resonance frequency ( $\omega \approx \omega_0/2$ ) and the component of the vibration at  $\omega_0$  is then measured. One thus has access to the value of the direct voltage  $V$ , which includes the direct voltage applied to the tip, but also certain electrical characteristics of the sample. On a metal, the tip-sample contact potential contributes to this potential difference. By this method, one can produce contact potential images. Due to their high sensitivity towards contamination



**Fig. 4.17.** (a) Topographic image of three carbon nanotubes connected between two gold electrodes on an oxide surface. (b) EFM image before cutting when a potential difference is applied across the electrodes. (c) EFM image after cutting the tubes by applying a voltage step at the points indicated by *arrows* in the first two images. In these images, lighter regions represent higher voltages [7]

layers, such images can prove useful in fields other than microelectronics. On an insulating layer, one can also detect the field created by localised charges. Triboelectric phenomena have also been investigated by this technique. On a semiconductor, the voltage will depend in a nonlinear way on the potential difference applied between the tip and sample, whereupon it may reveal doping levels in the material.

Let us mention another method that serves to measure the doping characteristics of semiconductors: scanning capacitance microscopy (SCM). In this technique, the tip is held in contact with the semiconductor sample and the charge of the metal tip–semiconductor capacitance across the passivation layer is modulated at several kHz by applying an electric field. (The depth of the depletion layer is modified.) A UHF resonance circuit is connected to this condenser in order to measure the small variations in the capacitance, with a sensitivity as high as  $10^{-18}$  farad.

Yet another technique for electrical analysis consists in measuring the contact resistance of the tip in scanning spreading resistance microscopy (SSRM).

## 4.8 Measuring Mechanical Properties

AFM can be used to describe the mechanical properties of the tip-sample contact. From here, one can then identify the Young's modulus, the hardness and local tribological properties of the surface. The corresponding methods are described in this section. Many techniques employed on a macroscopic scale can thus be implemented locally with an AFM, e.g., studies of adhesion or scratch tests, which are not discussed here.

### 4.8.1 Nanoindentation

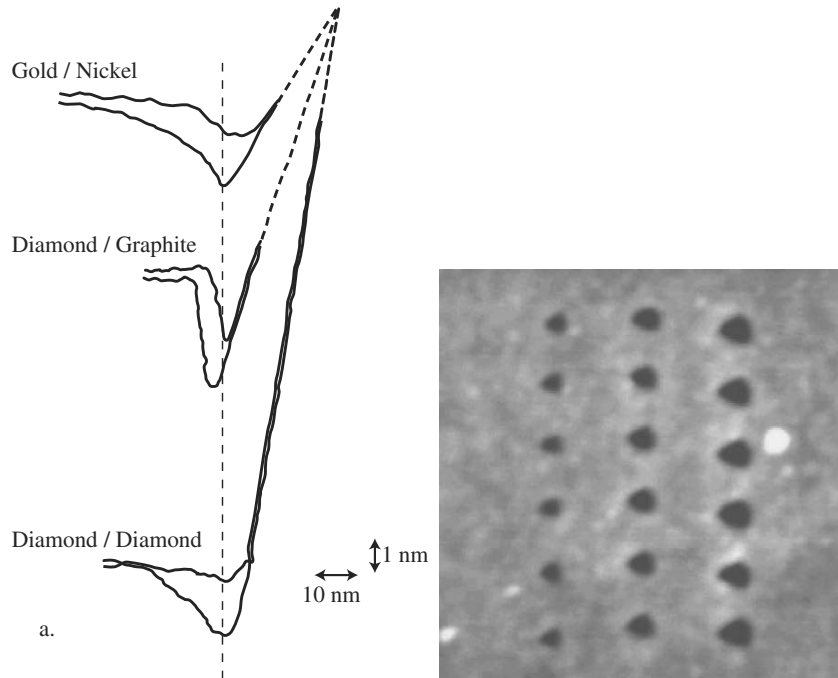
With its force sensor and displacement control system, the AFM can be used as a nanoindenter. Force curves are considered here as instrumented indentation curves. For hardness measurements, very stiff cantilevers and hard tips are generally used, the tip being made of diamond, or coated with diamond.

The geometry of the system leads to some difficulties in interpreting the data. The cantilever axis is tilted with respect to the plane of the surface. Hence a vertical displacement is always accompanied by a lateral force on the contact. To rectify this problem, the sample can be shifted laterally during the experiment in such a way as to cancel the displacement of geometrical origin.

The results obtained by this approach are often rather qualitative. Figure 4.18a shows experimental curves for different tip-sample pairs. When a quantitative interpretation is required, it must be obtained by modelling the contact. For plastic deformations, one can also measure the image size of the imprint left by the indenter. Figure 4.18b thus shows two groups of indentations obtained with loads of about 15, 20 and 25  $\mu\text{N}$  on a 15-nm diamond film. The deepest indentation is less than 10 nm.

### 4.8.2 Measuring Contact Stiffness

With the help of the force curves, one can also measure the elastic or viscoelastic indentation of the tip in materials whose plastic threshold is not too low, e.g., elastomers. Although the method has sometimes been used for highly deformable materials, there are several drawbacks. To achieve an acceptable level of sensitivity, the experiment must be carried out with a cantilever stiffness close to the contact stiffness. But under such conditions, the measurement is made neither by displacement nor by imposed force, and this complicates the analysis in the case of viscoelastic samples. Moreover, as in the case of indentation, the lateral component of the force on the contact, arising due to the geometry of the experiment, contributes to the recorded deflection. Finally, adhesion phenomena, which are often significant on the nanometric scale and bring in their own time constants, are liable to push the tip-sample system into a non-equilibrium state during the force curve measurement. It is then impossible to analyse results in terms of Young's modulus.

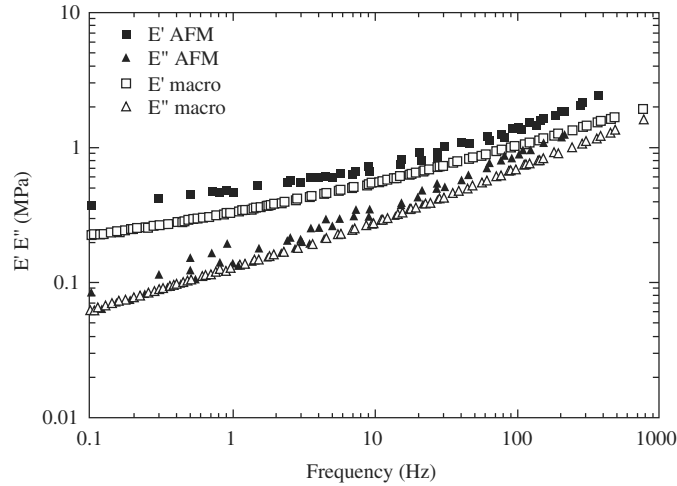


**Fig. 4.18.** Indentation experiments. *Left:* Force curves obtained with different tips on nickel, graphite and diamond [8]. *Right:* Indentations obtained with loads of about 15, 20 and 25  $\mu\text{N}$  on a 15-nm diamond film. The imprints are all less than 10 nm deep (see [www.veeco.com](http://www.veeco.com))

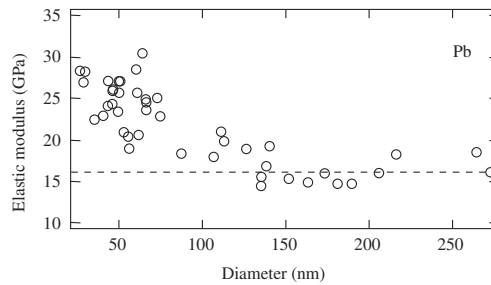
It is better to work at a fixed point, without scanning, by laterally modifying the sample position. Figure 4.19 shows the results obtained on an elastomer in viscoelastic phase. It has been possible to determine the Young's modulus of the material from measurements of the lateral stiffness of the contact under sinusoidal excitation. The results are compared here with a macroscopic measurement of this property.

#### 4.8.3 Contact Resonance Frequency

A useful alternative to these methods consists in measuring the resonance frequency of the cantilever when the tip is actually in contact. Indeed, the beam vibrates in a characteristic way for the boundary condition imposed by contact. By virtue of extremely accurate frequency measurements, the amplitude of the deflections can be maintained at a very low level. In this way, one can greatly reduce the lateral effects mentioned above. This technique has been successfully used to measure the Young's modulus of metallic nanotubes with diameters of a few tens to a few hundred nanometers (see Fig. 4.20).



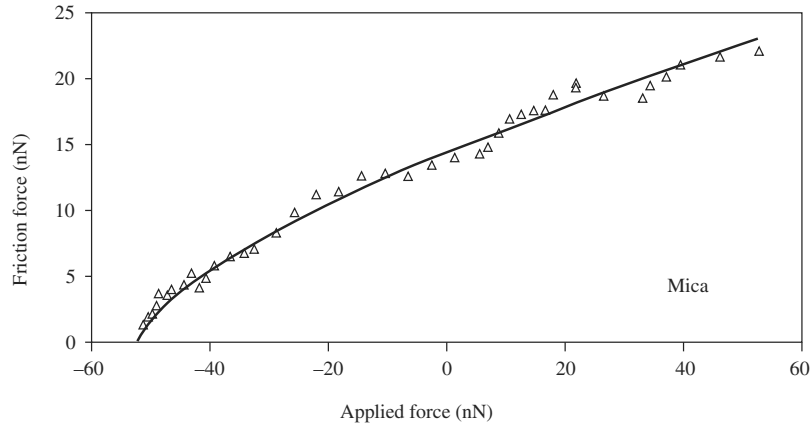
**Fig. 4.19.** Complex viscoelastic modulus of a latex film deduced from measurements in the static friction regime. Values are compared with those obtained on a macroscopic scale. (Courtesy of C. Basire and C. Frétiigny)



**Fig. 4.20.** Young's moduli of lead nanotubes as determined by measuring the resonance frequency of the cantilever when the tip is in contact. The macroscopic modulus, indicated by the *horizontal dashed line*, corresponds for the larger diameters, but not for smaller ones. In fact, the modulus increases significantly as the size decreases. The nanotubes are scattered on a porous membrane and measurements are carried out on structures that are caught in the pores. (Courtesy of S. Cuenot, C. Frétiigny, S. Dumoustier-Champagne and B. Nysten)

#### 4.8.4 Friction Forces

We have seen that the friction force recorded through the torsion of the cantilever during lateral scanning can be considered as a simple way of diagnosing the distribution of physicochemical parameters on a surface. Moreover, AFM can tackle more fundamental questions concerning friction. In particular, the absence of roughness effects on the scale of the AFM tip means that one can seek the underlying microscopic laws.



**Fig. 4.21.** Friction law measured on a mica surface. Results have been fitted to a  $2/3$  power law in the force [9]

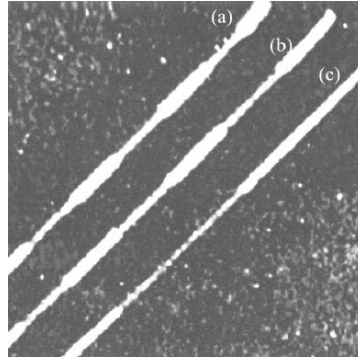
In some situations, a  $2/3$  power law is measured with respect to the load (shifted by an adhesive force). Such a case is represented by the data shown in Fig. 4.21 for a mica sample [9]. This relationship suggests a constant shear rate at the interface between the sample and the spherical tip, where the size of the contact zone varies according to the DMT theory.

## 4.9 Applications in Nanotechnology

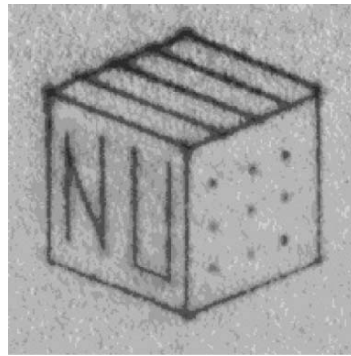
With a view to developing the various branches of nanotechnology, many research teams are engaged in the search for experimental protocols to etch, to write or to manipulate molecules and small structures on surfaces. The contact and tapping modes have been successfully used in a wide range of experimental contexts. A very complete review of these nanofabrication techniques can be found in [16]. There are many examples: displacement of carbon nanotubes or  $C_{60}$  molecules, local evaporation of gold deposited as a thin film on the tip, anodic oxidation of a surface in a damp atmosphere, wear, species desorption by heating, to name but a few. The tip can function as an electrode in an electrochemical reaction, a local heat source, and so on. A pierced cantilever can behave as a moving mask for an atomic beam, the so-called nanostencil. In the present section, we shall restrict ourselves to three examples to illustrate this diversity.

By oxidising silicon in air under the AFM tip in intermittent contact, one can produce linear oxide structures of width 15 nm and height about 1 nm. By chemical etching, silicon nanowires can then be fabricated. Figure 4.22 shows a  $0.7 \times 0.7 \mu\text{m}^2$  image of oxide lines with modulated width.

Another form of writing is obtained by using the AFM tip like the nib of a fountain pen. After dipping the tip into a thiol solution, the gold sample-tip



**Fig. 4.22.**  $0.7 \times 0.7 \mu\text{m}^2$  image of silicon oxide nanowires obtained by chemical etching under the AFM tip [10]

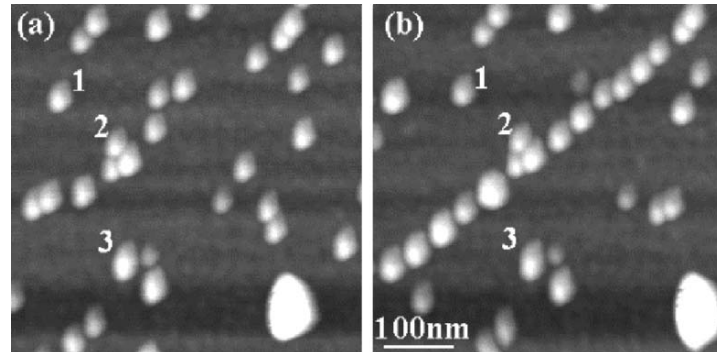


**Fig. 4.23.** Drawing obtained by so-called dip-pen lithography, in which thiol molecules are deposited on a gold surface. The line thickness is 30 nm [11]

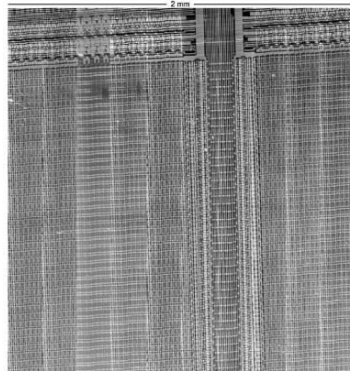
contact is used to transfer molecules through the water meniscus, whence they form a self-assembled lattice. A line thickness of 30 nm has been obtained by this very simple technique. Figure 4.23 shows an example.

Several investigations have shown that it is quite feasible to build ordered structures from a random deposit of small objects. The example in Fig. 4.24 (5-nm gold particles on silicon treated with poly-L-lysine) was obtained in non-contact mode by taking advantage of operating points in different regions of the tip-sample interaction potential.

One important aim in this kind of study is the reading or writing of information using AFM tips, with a view to increasing storage density. An IBM team in San Jose has perfected a micromechanical system on a rotating disk on which, in read-only, the tip recognises structures of 100 nm. The integration density is 10 gigabits per square centimeter, 100 times greater than a CD-ROM. Another device, still in the experimental phase, this time single-write and read, uses an AFM tip heated by electrical pulses. This creates dips in a



**Fig. 4.24.** Starting from a random deposit, 5-nm gold particles are displaced so as to form a straight line [12]



**Fig. 4.25.** Printed circuit obtained using 10 cantilevers in parallel ( $2 \times 2 \text{ mm}^2$ ) [13]

polymer layer which it can then interpret as bits in read mode. The density obtained here is 5 gigabits per square centimeter with a read rate of 10 Mb/s. The series write restriction when using an AFM tip makes it particularly slow and hence economically unviable. In order to go beyond the 200–300 Mb/s range of existing magnetic systems, research teams therefore build parallel arrays of cantilevers. A 2D device comprising  $32 \times 32$  tips over an area of  $3 \times 3 \text{ mm}^2$  has been achieved. If structures and spacings of the order of 40 nm can be produced, this should lead to speeds as great as a few hundred Mb/s.

The technology of cantilever beams acting in parallel has also been developed by Quate at Stanford University for the purposes of imaging or acting very quickly over large areas. Figure 4.25 shows a  $2 \times 2 \text{ mm}^2$  printed circuit (image 25 million pixels), obtained using 10 cantilevers in parallel for 30 min.

To end this section, we should mention the development of a force feedback nanomanipulator, which is a system coupling the microscope to a virtual reality interface in such a way as to give the user the feeling of actually being on



the sample surface, with a magnification scale factor of one million [15]. This tool may prove useful in the development of complex fabrication processes.

## 4.10 Conclusion

Force microscopy is now widely used for routine analysis in many research establishments. However, the AFM is far from being a simple imaging system. It can also be considered as the basic component of a wide range of instruments, with applications ranging from fundamental physics to nanolithography, and it is in this context that it plays such an important role in the development of nanotechnology. This technique is still moving forward rapidly. Apart from the examples chosen here to illustrate the main operating modes of the AFM, there are many other uses, each fulfilling specific requirements in fundamental or applied science.

## References

1. Garcia, R., and Perez, R.: *Surf. Sci. Rep.* **47**, 197–301 (2002)
2. Sounilhac, S.: *Doctoral Thesis, Paris XI* (1998)
3. Thurn-Albrecht, T.: *Phys. Rev. Lett.* **87**, 22 (2001)
4. Lantz, M.A., Hug, H.J., van Schendel, P.J.A., Hoffmann, R., Martin, S., Baratoff, A., Abdurixit, A., Güntherodt, H.-J., and Gerber, Ch.: *Phys. Rev. Lett.* **84**, 12 (2000)
5. Ducker, W.A., et al.: *Nature* **353** (1991)
6. Rief, M., et al.: *Science* **275**, 1295 (1997)
7. Park, J.-Y., et al.: *Appl. Phys. Lett.* **80**, 23 (2002)
8. Burnhamt, N.A., Colton, R.J., and Pollock, H.M.: *Nanotechnology* **4**, 64–80 (1993)
9. Pietremont, O., et al.: *Tribology Lett.* **7**, 213 (1999)
10. Legrand, B., Deresmes, D., and Stievenard, D.: *J. of Vac. Sci. Tech.* **20** (3), 862–870 (2002)
11. Piner, R.D., et al.: *Science* **283**, 661 (1999)
12. Ramachandran, T.R., Baur, C., Bugacov, A., Madhukar, A., Koel, B.E., Requicha, A., and Gazen, C.: *Nanotechnology* **9**, 3 (1998)
13. Minne, S.C., Adams, J.D., Yaralioglu, G., Manalis, S.R., Atalar, A., and Quate, C.F.: *Appl. Phys. Lett.* **73**, 12, 1742–1744 (1998)
14. Reiter, G., Castelein, G., Sommer, J.-U., Röttele, A., and Thurn-Albercht, T.: *Phys. Rev. Lett.* **87** (2001)
15. Falvo, M.R.: *Nature* **389**, 583 (1997)

## General References

16. Basski, A.A.: *Recent and Evolving Advanced Semiconductors and Organic Nanotechniques*, Part 3, ed. by H. Morkoc, Academic Press (2002)
17. Sarid, D.: *Scanning Force Microscopy with Applications to Electric, Magnetic and Atomic Forces*, Oxford University Press (1994)
18. Lange, D.: *Cantilever-Based CMOS Nano-Electro-Mechanical Systems: Atomic-Force Microscopy and Gas Sensing Applications*, Springer-Verlag (2002)
19. Meyer, E.: *Atomic Force Microscopy: Fundamentals to Most Advanced Applications*, Springer-Verlag, New York (2002)
20. Tsukruk, V. (Ed.): *Advances in Scanning Probe Microscopy* (Macromolecular Symposia 167), Wiley, New York (2001)
21. Soh, H.T., Guarini, K.W., Quate, C.F.: *Scanning Probe Lithography* (Microsystems, Vol. 7), Kluwer Academic Publishers (2001)
22. De Stefanis, A., Tomlinson, A.A.G.: *Scanning Probe Microscopies: From Surface Structure to Nano-Scale Engineering*, Trans Tech Publications (2001)
23. Bonnell, D.A. (Ed.): *Scanning Probe Microscopy and Spectroscopy: Theory, Techniques, and Applications*, Wiley, New York (2000)
24. Sakurai, T., Watanabe, Y. (Eds.): *Advances in Scanning Probe Microscopy* (Advances in Materials Research, 2), Springer-Verlag (2000)
25. Minne, S.C., Manalis, S.R., Quate, C.F.: *Bringing Scanning Probe Microscopy up to Speed* [Microsystems (Series), 3], Kluwer Academic Publishers (1999)
26. Concerning the development and applications of local probe microscopies: Bouhacina, T., Kopp-Marsaudon, S., Aimé, J.P.: *Spectra 2000 Analyse* (1998), Vol. 27, No. 203, pp. 11–20
27. Concerning near-field microscopies: Marsaudon, S., Bouhacina, T., Aimé, J.P.: review article in *Spectra Analyse*, Vol. 31, No. 225, pp. 15–25 and No. 227, pp. 13–21 (2002)
28. Concerning local probe microscopies: From atomic imaging to nanoscale spectroscopy: Nysten, B.: *Chimie Nouvelle* **18**, 3059–3070 (2000). Also available at [www.mapr.ucl.ac.be/nysten/SPMs.ChimNouv.pdf](http://www.mapr.ucl.ac.be/nysten/SPMs.ChimNouv.pdf)
29. Chemical and biological analysis using AFM: H. Takano, Kenseth, J.R., Wong, S.-S., O'Brien, J.C., Porter, M.D.: *Chem. Rev.* **99**, 2845–2890 (1999)
30. Near-field microscopes: Frétigny, C.: Belin (to be published 2005)
31. Websites of AFM manufacturers and suppliers of associated equipment provide technical information concerning operating principles and apparatus, spectacular examples of applications, and many links to research centers, journals, and so on. A complete list of sites can be found at the address [www.nanoworld.org/english/companies.htm](http://www.nanoworld.org/english/companies.htm)

## Near-Field Optics: From Experiment to Theory

C. Boccara and R. Carminati

The aim in this chapter is to provide the reader with an overview of near-field optics, covering the basic principles, experimental approach, and theory. We have chosen to base the discussion on familiar observational phenomena in order to explain why resolution in optics is not only limited, as is often believed, by the instrument (microscope, photographic objective, etc.) and the wavelength used, but also by the object itself. To go beyond classical resolution limits, one must enter the domain of near-field optics. There are various ways of gaining access to the electromagnetic field in the immediate vicinity of a sample surface. We shall estimate the orders of magnitude of the main physical effects encountered, emphasising the novel features that should be revealed by this new form of microscopy. Once we have prepared the ground, we proceed with a predominantly theoretical account of the basic concepts underlying near-field optics in terms of the angular spectrum of plane waves and radiation by simple or more complex structures.

### 5.1 Basic Ideas and the Nature of the Problem

#### 5.1.1 Resolution, Near Field and Far Field

Let us consider some simple examples:

- Consider a diffraction grating with grating interval  $p$ , ruled alternately with opaque and transparent lines. Suppose that it is illuminated at normal incidence by a monochromatic plane wave of wavelength  $\lambda$ . This grating diffracts orders  $k = \pm 1, \pm 2$ , etc., which emerge at angles  $i'$  such that  $\sin i' = k\lambda/p$ . If the grating interval  $p$  is less than  $\lambda$ , only the zero order will be produced. Does this mean that the information indicating that the object is periodic, which is contained in the orders other than zero, has been lost? This information exists at the surface of the grating, e.g., through the boundary conditions we impose for the field of the wave, but

if  $\lambda > p$ , it does not propagate, so to speak, and remains localised close to the surface in the form of so-called evanescent waves.

- Optical instruments are traditionally considered as a kind of filter for the spatial frequencies contained within the object. This filtering appears in the Fourier plane when we close the pupil of a  $4f$  setup, for example [1]. A diffraction calculation shows that, for a microscope objective, the resolution, i.e., the smallest discernible separation between two point objects in the absence of aberrations, is given by

$$\rho = \frac{1.22\lambda}{2n \sin u} ,$$

where  $n$  and  $u$  are the refractive index and aperture in the object space, respectively. It follows that, for a visible wavelength ( $\lambda = 0.5 \mu\text{m}$ ) and an object immersed in oil ( $n = 1.5$ ),

$$\rho_{\min} \approx 0.2 \mu\text{m} .$$

If the above grating were illuminated obliquely, at an almost grazing angle, and if it were immersed in a medium with refractive index  $n$ , the condition for obtaining a diffracted order other than the zero order would be  $p > \lambda/2n = 0.17 \mu\text{m}$ , which leads to the same orders of magnitude.

- A plane wave which illuminates a slit of width  $a \gg \lambda$  emerges almost parallel to its direction of incidence. If  $a$  is then decreased, the beam diverges, i.e., the projection of the wave vector  $\mathbf{k}$  (with  $k = 2\pi/\lambda$ ) onto the plane of the slit increases. This projection reaches the value  $2\pi/\lambda$  when  $a = \lambda/2$ , whereupon the light fills the whole half-space. One may wonder what happens beyond this point. If we continue to narrow the slit, no further information such as might inform us as to the slit size, for example, can reach us through the form of the angular distribution. However, we will show later that the component of the wave vector in the plane of the slit continues to increase, whilst the corresponding waves simply do not propagate, but remain evanescent.

We conclude that, if we wish to obtain an ‘optical’ image, i.e., with electromagnetic waves of wavelength  $\lambda$ , with better resolution than  $\lambda/2n$ , we must somehow pick up these evanescent waves directly as they manifest themselves on the surface of the object.

### 5.1.2 Brief History of Near-Field Methods

The ideas introduced briefly in the last section were quite clear to physicists in the 1930s, as attested by the correspondence<sup>1</sup> between Synge and Einstein.

<sup>1</sup> This correspondence was brought to our attention by Daniel Courjon, University of Franche Comté, Besançon, France, one of the pioneers of near-field optical microscopy.

Synge suggested exploring the near field by means of a scattering nanoparticle, viz., a gold particle, placed under a microscope slide and used to scan the sample surface. The role of the scattering particle was to render the evanescent fields ‘propagative’. This is the idea underlying the second method that we shall describe, known as apertureless near-field optical microscopy.

Einstein found this approach rather delicate to implement and suggested using a nano-aperture in a silver film deposited on glass. This aperture could then act as a nanoscale light source for scanning the sample surface.

Finally, Synge had the idea of coating a cone with metal, and making a nano-aperture in the apex of the cone, since this would be easier to bring into proximity with the sample. Today, this is the predominant way of doing things, in particular, in scanning near-field optical microscopy (SNOM).

### 5.1.3 Near-Field Optical Microscopy: For What Purpose?

An optical microscope used in the far field is an extremely effective tool, operating close to the physical limits imposed by diffraction and easy to implement. An optical microscope used in the near field is a priori a much more delicate instrument to put into practice and must be operated judiciously, bearing in mind the following points:

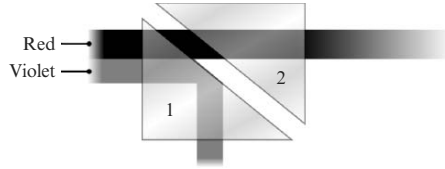
- If one is only concerned with topographical contrast on the scale of a few nanometers or less, the STM methods discussed in Chap. 3 for conducting samples, or the AFM methods discussed in Chap. 4 for insulating samples are without doubt more appropriate.
- Near-field optical microscopy should make it possible to exhibit optical contrasts on a scale well below that of the wavelength.

The most frequently encountered optical contrasts can be attributed to the following items:

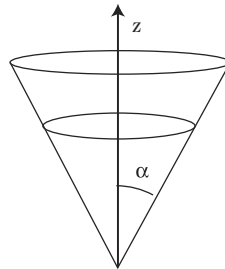
- The dielectric constant  $\epsilon_r$  or the refractive index  $n$ , where  $\epsilon_r = n^2$ . If the index is a complex number of the form  $\tilde{n} = n + ik$ , the contrast will spring from phase and amplitude effects in the wave, via  $n$  and  $k$ , respectively.
- Local fluorescence. Fluorescent molecules and nanostructures receive a great deal of attention in this field.
- Magneto-optical response. This arises from anisotropies in the index  $n$  for two orthogonal polarisations (circular or linear).
- Local photochemical reactions, as used in lithography.

## 5.2 Photon Scanning Tunneling Microscope (PSTM)

This approach uses a tapered monomode optical fibre to probe the evanescent field produced when evanescent waves are created by total reflection. We shall now go through the physical phenomena involved one by one.



**Fig. 5.1.** Reflection is total for violet but frustrated for red, if the separation between the two prisms is chosen suitably



**Fig. 5.2.** Optical fibre with tip tapered into a conical shape having semi-angle  $\alpha$  at the apex. It is polarised by the evanescent field surrounding it

### 5.2.1 Frustration of Evanescent Fields

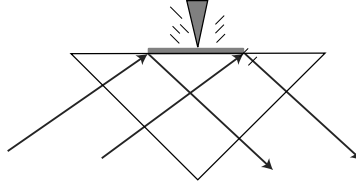
By analogy with the tunnel effect in quantum mechanics (see Chap. 3), the evanescent wave existing at the surface of an object when there is total reflection can be frustrated by the presence of a third medium (the first being the object and the second the vacuum or air), i.e., this third medium can prevent total reflection from actually occurring.

The experiment depicted in Fig. 5.1 shows that, in the case of a totally reflecting prism (1), the presence of the second prism (2) frustrates reflection for the longest wavelengths (red), in particular, for  $\lambda > 2\pi e$ . More precisely, we can write down the boundary conditions and find the field distribution in each medium.

Whilst this 1D model can be used to quantify the frustration, it gives us no idea of the resolution that could be obtained with a conical tip brought in to probe the evanescent field locally (see Fig. 5.2). Indeed, this is the setup used in the photon scanning tunneling microscope (PSTM), initially suggested at two French research centers at Besançon and Dijon [6, 7].

### 5.2.2 PSTM Probe in an Evanescent Field: Scattering Model

The 1D frustration model is no longer suitable on the scale of the finely tapered fibre tip. The very end of the tip is considered here as a small scattering object of dimensions  $\ll \lambda$ . There is no simple way to calculate the polarisation of a conical object in a non-uniform field. However, if the dielectric constant is



**Fig. 5.3.** The tapered dielectric fibre radiates the local evanescent field close to the fundamental mode of the fibre

close to unity, the depolarising fields may be neglected and the polarisation summed over slices of the object taken perpendicular to the  $z$  axis, in order to obtain a reasonable estimate. The magnitude of the polarisation vector is thus found to be  $P = \varepsilon_0(\varepsilon_r - 1)EV$ , where  $E$  is the applied field and  $V$  the volume affected by the field  $E$ .

For example, for a sphere, this result should be compared with the standard formula

$$P = 3\varepsilon_0 \frac{\varepsilon_r - 1}{\varepsilon_r + 2} EV .$$

Applying this to silicon, for which  $\varepsilon_r = n^2 \approx 2$ , the formula gives the right order of magnitude, since the error is around 25%.

The evanescent field falls off exponentially as one moves away from the surface. Let  $S$  be the characteristic length with which the field drops off. We may now calculate

$$P = \int_0^Z dv(z) \varepsilon_0 (n^2 - 1) E_0 e^{-z/S} , \quad \text{where } dv(z) = \pi r^2 dz = \pi z^2 \tan^2 \alpha dz .$$

Integrating by parts, we find that what one may call the useful volume is  $2S^3 \tan^2 \alpha$ . (We recall that for an evanescent wave created by total reflection beyond the critical angle,  $S \approx \lambda/2\pi$ .) This gives an idea of the useful dimensions of the probe. The mean volume of the scattering probe can be found by putting  $\alpha = 1/10$  rad and  $\lambda = 0.6 \mu\text{m}$ , which gives a value of the order of  $20 \times 10^{-6} \mu\text{m}^3$ . This confirms the useful dimensions expected for such probes, of the order of ten nanometers.

We may now calculate the power radiated by this dipole, i.e., the energy collected for the measurement:

$$W = \frac{1}{4\pi\varepsilon_0} \frac{\omega^4 P^2}{3c^2} , \quad \text{where } P = 2\pi \tan^2 \alpha S^3 .$$

In practice, not all the scattered energy is collected, but only that part of it which is coupled to the fundamental mode of the fibre at the end where the detector is located (see Fig. 5.3).

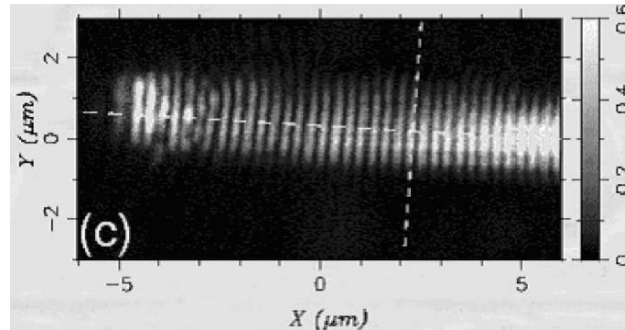


Fig. 5.4. PSTM image of a guide 2.5  $\mu\text{m}$  wide, excited by a surface plasmon [12]

### 5.2.3 Applications of PSTM

Since the PSTM was an early introduction to the optical near-field scene, a great deal of work was carried out to test this experimental approach, particularly with regard to its resolution. These tests often sought to identify topographical structure, such as steps made from the same material as the totally reflecting prism, with no other contrast than that associated with phase offset.

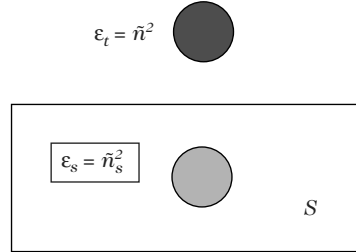
It would seem that the real interest of PSTM is in investigating structures of a kind that only generate evanescent waves. Wave guides or more sophisticated components of planar integrated optics provide a striking example.

More recently, new structures have been developed, guiding light by means of metallic dots much smaller than the wavelength of the light (plasmon guides). As the PSTM has very little perturbative effect on propagation, it is destined to become the preferred tool for measuring the field distributions in such wave guides. Figure 5.4 illustrates how a guiding effect can be probed by the PSTM [12].

## 5.3 Apertureless Near-Field Microscope

We have just seen how the end of the dielectric probe can sample the evanescent field locally over a volume much smaller than the light wavelength. In the case of the PSTM, the localisation of the response is due partly to the nature of the evanescent field, which falls off as one moves away from the surface, and partly to the small probe size. We now turn to another object capable of scattering a near field into the far field. However, this probe, which is generally metallic, has two very useful properties: the field probed under the tip is highly localised and there is very efficient scattering.





**Fig. 5.5.** Simplified view of a microscope with scattering metallic tip. We consider here the influence of the medium with dielectric constant  $\varepsilon_s = \tilde{n}_s^2$  on radiation from the spherical probe

### 5.3.1 Nano-Antenna Radiating to the Far Field

In this approach, the apex of a tip, which is either metallic or uses a high-index dielectric, radiates the local field into a light collector (a mirror or wide-aperture microscope objective) which focuses it onto a sensor. This device is similar to the PSTM in some ways. The difference comes from the fact that apertureless setups rarely use total reflection, and also that the tips play an active role in localising the field and scattering it efficiently. On the other hand, the presence of a metal can perturb the local distribution of the field. Only when the tip apex occupies a very small volume, so that the perturbation is reduced accordingly, can the tip be considered as a passive probe.

### 5.3.2 Source of Contrast: Scattering Sphere Model

The aim here is to understand the influence of the local dielectric constant on the probe which scatters the electromagnetic field located at the sample surface (evanescent and propagative waves) into the far field. Although the use of a sphere to represent the tip is no doubt restrictive, it nevertheless gives a good idea of the physical phenomena involved.

The probe considered here is a scattering sphere with radius  $a$  and dielectric constant  $\varepsilon_t = \tilde{n}^2$ , which is easier to treat than the metal tip itself, in the vicinity of a medium with dielectric constant  $\varepsilon_s = \tilde{n}_s^2$  (see Fig. 5.5). When the medium is not present, we use the electrostatic approximation wherein the field  $\mathbf{E}$  of the electromagnetic wave polarises the sphere and gives rise to a dipole  $\mathbf{p}$  such that

$$\mathbf{p} = (\varepsilon_t - 1)\varepsilon_0 \mathbf{E}V = \alpha \varepsilon_0 \mathbf{E} ,$$

with

$$\alpha = 4\pi a^3 \frac{\varepsilon_t - 1}{\varepsilon_t + 2} .$$

The presence of the second medium modifies the field created by this dipole in two ways (see the chapter<sup>2</sup> on electrostatic images in [13]):

- There is an image dipole

$$\mathbf{p}' = -\mathbf{p} \frac{\varepsilon_s - 1}{\varepsilon_s + 1} .$$

- $\mathbf{p}$  becomes

$$\mathbf{p} \frac{2\varepsilon_s}{\varepsilon_s + 1} .$$

Consider the example of the dipole induced on the sphere of radius  $a$  located at distance  $d$  from the surface by the field  $\mathbf{E}$  of an electromagnetic wave polarised normally to the sample surface. The field created at distance  $r$  by the image dipole (field component normal to the surface) is

$$E' = \frac{1}{4\pi\varepsilon_0} \frac{2p'}{r^3} = \frac{1}{2\pi\varepsilon_0} \frac{p'}{2^3(a+d)^3} ,$$

or

$$p' = -\alpha\varepsilon_0 E \frac{\varepsilon_s - 1}{\varepsilon_s + 1} , \quad \text{since } p' = -\alpha\varepsilon_0 E \frac{\varepsilon_s - 1}{\varepsilon_s + 1} ,$$

whereupon

$$E' = -\frac{\alpha}{2\pi} \frac{E}{2^3(a+d)^3} \frac{\varepsilon_s - 1}{\varepsilon_s + 1} .$$

The total field polarising the sphere is

$$E + E' = E \left[ 1 - \frac{\alpha\beta}{16\pi(a+d)^3} \right] , \quad \text{where } \beta = \frac{\varepsilon_s - 1}{\varepsilon_s + 1} .$$

Finally, the dipole  $\mathbf{p}$  is itself modified to

$$p'' = p \frac{2\varepsilon_s}{\varepsilon_s + 1} = p(\beta + 1) .$$

It follows that

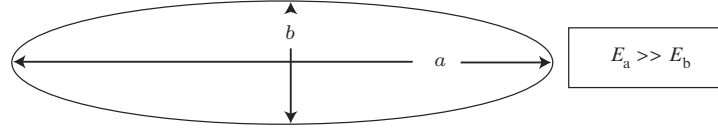
$$\alpha_{\perp\text{effective}} = \alpha \frac{\beta + 1}{1 - \frac{\alpha\beta}{16\pi(a+d)^3}} .$$

In order to establish the influence of the medium on the radiation from the sphere, it suffices to calculate the scattered energy, or scattering cross-section

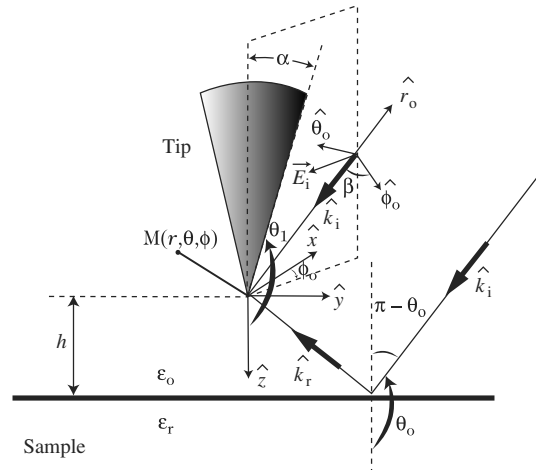
$$C_{\text{scatt}} = \frac{k^4 |\alpha_{\perp\text{eff}}|^2}{6\pi} ,$$

in the far field as a function of the distance  $d$  and the optical characteristics  $\varepsilon_r = \tilde{n}^2$  of the various media.

<sup>2</sup> What Jackson does for a charged particle can be applied to a dipole in a linear, homogeneous and isotropic medium.



**Fig. 5.6.** In electrostatics, the field is much stronger along the major axis of a metal ellipsoid than along its minor axis. This is the sharp-point effect



**Fig. 5.7.** A plane electromagnetic wave illuminates a metal tip either directly or after reflection from a sample. We seek the field distribution close to the tip

**5.3.3 Sharp-Point Effect. Tip Resolution and Efficiency**

The above calculation does not take into account the fact that, in an apertureless SNOM setup, the scattering object is not a sphere but a tip (often metallic). Now, in the near field, i.e., at distances  $\ll \lambda$  from the object, a quasi-electrostatic approximation can be made for electromagnetic phenomena. As a result, any sharply pointed object will see a greater field than an object of lower curvature. For example, for a metal ellipsoid of revolution with principal radii  $a$  and  $b$ , one has  $E_a/E_b = a/b$  (see Fig. 5.6).

We now consider the case of a metal scattering tip and show that, if this conical tip is irradiated by a plane wave, we will find a field localised under the apex which is significantly enhanced compared with the incident field (see Fig. 5.7).

**5.3.4 Field Enhancement Near a Metal Tip**

When the cone is illuminated by this wave, the expression for the total diffracted field at an observation point in the near field ( $kr \ll 1$ ) is given approximately by a product of three terms. The first depends on the tip geometry,

the second on the observation point, and the third on the incident wave [9,10]:

$$E(\theta_0) \approx \left\{ \frac{i \exp\left(-\frac{1}{2}p_1\pi\right)}{\sin \theta_1} \frac{\sqrt{\pi}}{2^{1/p_1} \Gamma\left[p_1 + \frac{1}{2} \frac{1}{P_{p_1}^1}(\cos \theta_1) \frac{\partial}{\partial p_1} P_{p_1}(\cos \theta_1)\right]} \right\} \\ \times \left[ (kr)^{1/p_1} \left( \hat{r} + \frac{\hat{\theta}}{p_1} \frac{\partial}{\partial \theta} \right) P_{p_1}(\cos \theta) \right] \times [P_{p_1}^1(\cos \theta) \sin \beta] , \quad (5.1)$$

where  $P_{p_1}^m$  are Legendre polynomials of order  $m$  and degree  $p_1$ . These functions are brought into the expression by the expansion of the electric potential  $V$  in spherical harmonics. In the near field of the cone, this potential can be written

$$V(r, \theta, \varphi) = \sum_{m=0}^{\infty} \sum_{p_1} r^{p_1} P_{p_1}^m(\cos \theta) (A_{mp_1} \sin m\varphi + B_{mp_1} \cos m\varphi) .$$

The value of  $p_1$  is chosen so that this potential vanishes on the cone, i.e.,  $P_{p_1}^m(\cos \theta_1) = 0$ .

For small angles, the value of  $p_1$  is related to  $\alpha$  by the approximation

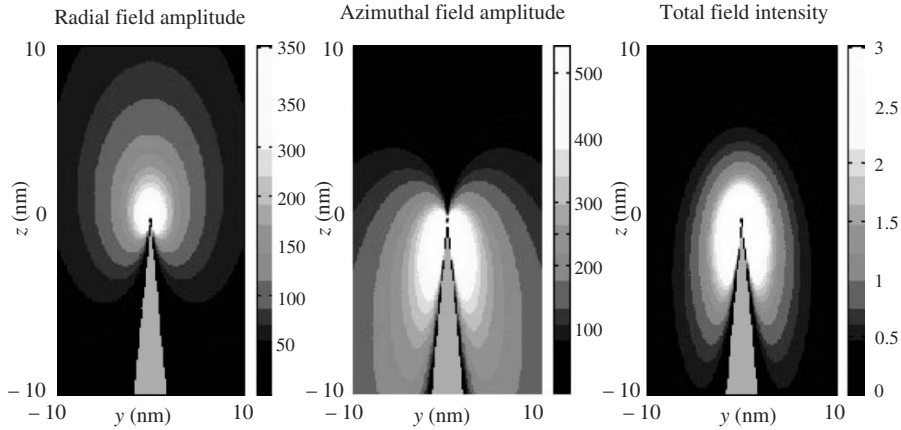
$$p_1 \approx \frac{1}{2 \ln(2/\alpha)} , \quad (5.2)$$

which can be used for angles  $\alpha$  not exceeding  $20^\circ$ . For this particular value of  $\alpha$ , the approximation gives 0.286 for  $p_1$ , whereas the correct value is 0.275, i.e., an error of the order of 4%. In fact,  $p_1$  tends to zero for small  $\alpha$ , whereupon the Legendre functions and their derivatives in (5.1) can be written in the asymptotic forms

$$P_{p_1}(\cos \theta) \approx 1 + 2p_1 \ln[\cos(\theta/2)] , \\ P_{p_1}^1(\cos \theta) \approx p_1 \tan(\theta/2) , \\ \frac{\partial}{\partial \theta} P_{p_1}(\cos \theta) \approx -p_1 \tan(\theta/2) , \\ \frac{\partial}{\partial p_1} P_{p_1}(\cos \theta) \approx 2 \ln[\cos(\theta/2)] . \quad (5.3)$$

Examination of fabricated tips by scanning electron microscope (SEM) shows that they have very low opening angles, never greater than  $5^\circ$ . Equation (5.1) and the approximations (5.3) can thus be used to calculate the intensity of the near field diffracted by this type of tip.

Figure 5.8 shows the spatial distribution in the incident plane of the intensity of the field diffracted by the tip, together with the amplitudes of its radial and azimuthal components. The incident wave has polarisation  $p$  (angle  $\beta = \pi/2$ ) and wavelength  $\lambda = 10 \mu\text{m}$ . The angle of incidence is  $\theta_0 = 100^\circ$  and



**Fig. 5.8.** Intensity of the electric field and its components near a metal tip

the opening angle is  $\alpha = 3^\circ$ . The extent of the intensity calculation is limited by the condition  $kr \ll 1$ . For  $r = 10$  nm, the product  $kr \sim 0.02 < 1$ .

The result of a numerical computation based on the above expressions shows a significant enhancement of the field amplitude directly beneath the tip, i.e., for  $\theta = 0$ . This enhancement is closely confined to within a few nanometers of the tip apex. The contribution of the radial component dominates directly beneath the tip, whereas around the edges of the cone ( $\theta \sim \theta_1$ ), the amplitude of the azimuthal component is predominant.

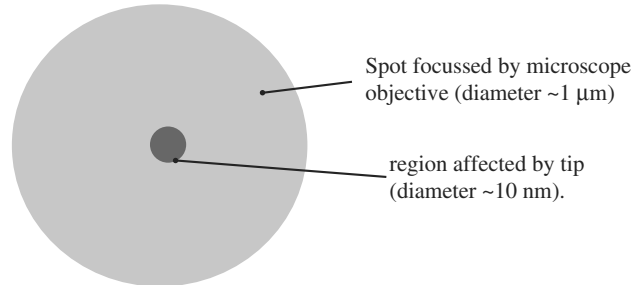
As we saw in the sphere model, the presence of the sample modifies the radiation from the metal tip. The theory of dielectric images can still be applied to determine the influence of the local dielectric constant, for example.

### 5.3.5 Apertureless SNOM: Typical SNOM Setup

Apertureless near-field microscopes have been used in different ways. They may operate by reflection, using the same objective to focus light on a spot with the same size as the diffraction pattern and to collect light reflected by the sample and scattered by the tip, which interferes on the detector (see Fig. 5.9). Note that this interference amplifies the weak signal of the near field scattered by the tip, because the specularly reflected flux is greater.

If the sample is transparent, the microscope may also operate by direct transmission of a focussed light beam, or by total internal reflection, as in the PSTM.

Finally, the surface can be illuminated by an almost-grazing wave. This configuration has been used to work in the infrared at  $10\ \mu\text{m}$ , with a  $\text{CO}_2$  laser as source. (The resolution of around 10 nm is still imposed by the tip size and not by the wavelength  $\lambda$ .) It is useful to describe this setup, illustrated in



**Fig. 5.9.** In an apertureless near-field microscope, the tip usually scatters a small part of the field obtained by focussing a laser with a microscope objective

Fig. 5.10, because, despite the specificity of certain of its features, it exhibits a number of others that are common to a great many SNOM or PSTM setups.

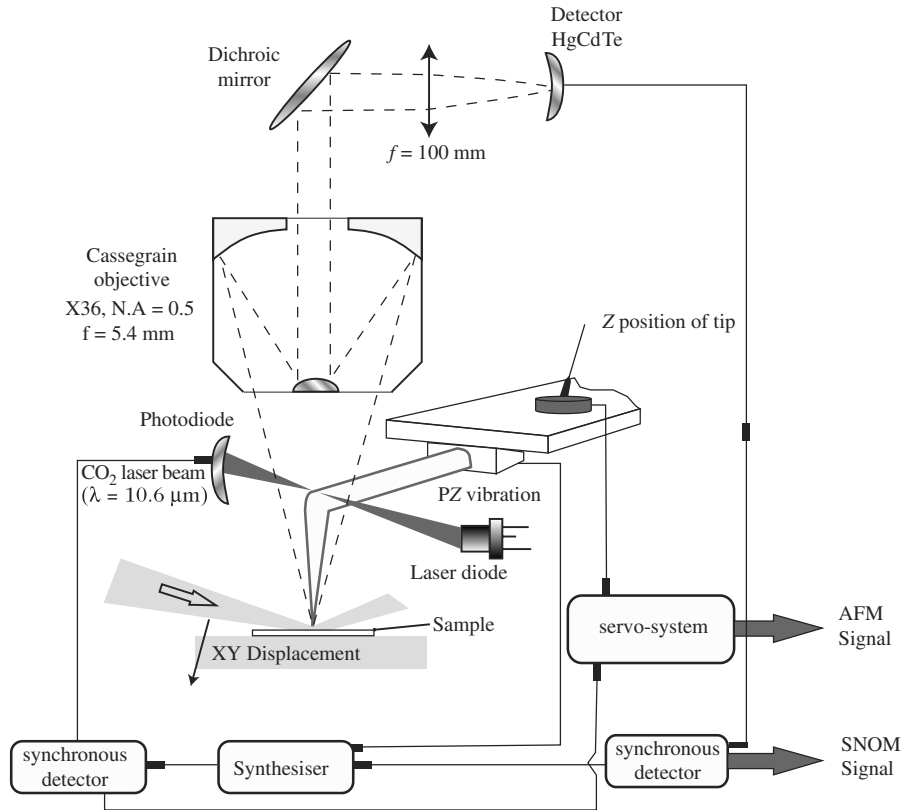
During the relative displacement of the tip and sample, the probe, in this case a metal tip, must follow the surface in order to describe the local fields in well-defined conditions and at the same time reveal the topography. The SNOM thus comprises two distinct but coupled parts:

- An atomic force microscope (AFM) with a feedback loop to maintain a constant tip-sample separation. Here the tip oscillates, being in the tapping mode, and the oscillation amplitude is measured by an optical system and synchronous detection so that it can be held constant during scanning. Other setups use shear-force mode.
- An optical part, in this case a microscope with a mirror objective due to the wavelength used. The scattered signal is modulated by the oscillation of the tip, which moves from the very near field to a slightly more distant zone (typically, ten to a hundred nanometers). Once again, the detector signal is demodulated by synchronous detection. The optical microscope allows one to choose the part of the field to be explored with the tip.

The modulation serves to separate the far-field signal from the near-field signal and to filter the useful signal from the noise. Frequencies used vary from a few kHz to several tens of kHz, so that one can reach the shot noise associated with scattered photons even if there are not many of them. This is due to a homodyne effect which occurs on the detector through interference of the near and far fields (reflected, scattered or transmitted by the sample).

Returning to the setup shown in Fig. 5.10, the IR detector is a CdHgTe photodiode, cooled with liquid nitrogen and particularly sensitive at 10  $\mu\text{m}$ . Any kind of detector, e.g., photodiode, cascade photodiode, photomultiplier, etc., can be used, depending on the application.

As far as the light source is concerned, lasers with different wavelengths (UV, visible, IR) have been used, and so have incoherent sources such as arc lamps coupled to a monochromator, thus confirming the effectiveness of the



**Fig. 5.10.** A scanning near-field optical microscope (SNOM) always involves a parallel function as an atomic force microscope (AFM)

method. It is even possible to use the radiation emitted by the sample itself, revealing the highly specific properties of thermal emission in the near field.

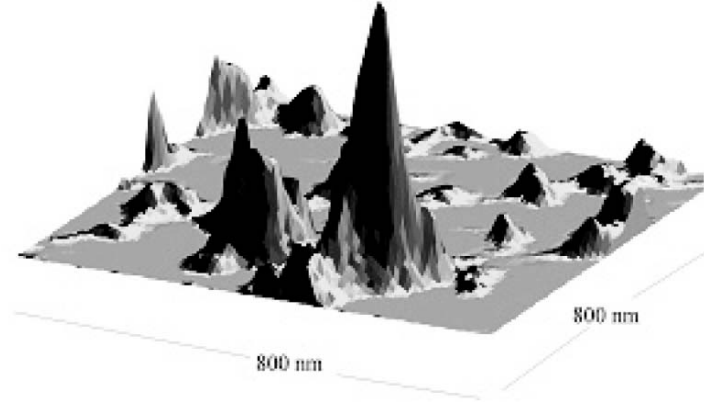
As an example, Fig. 5.11 shows how the method can probe fields on a very small length scale, typical resolution being around ten nanometers.

## 5.4 Aperture SNOM

In this paragraph, we shall discuss the modern version of Synge's original idea [19], which consisted in using a small aperture in a metal screen as a nanoscale light source.

### 5.4.1 Metal-Coated Fibre

As noted earlier, the idea here is to 'force' the light to go through a small hole with dimensions much smaller than its wavelength. One generally uses a



**Fig. 5.11.** Electromagnetic field distribution at a glass surface coated with gold particles at the percolation threshold and illuminated at 800 nm. Apertureless SNOM reveals peaks in the field with spatial extent varying between ten and a few hundred nanometers. (Image courtesy of Samuel Grésillon)

tapered optical fibre that has been coated with metal at the end, except for the hole, to compel the field to remain within the fibre right up to the hole. Recall that for a non-metallised dielectric fibre, the fundamental mode always propagates, whatever the thickness of the fibre core, the main part of the field being located outside the core in the form of an external evanescent field [18].

The hole can be used as a nanoscale light source, by transmitting the light emitted by a source placed at the other end of the fibre, or as a nanoscale detector, by transmitting the light along to a detector located at the other end of the fibre. However, the amount of light which manages to emerge from the fibre, or to reach the detector going the other way from the tip end, is extremely small, as we shall see.

#### 5.4.2 Energy Transmission in a Tapered Metal-Coated Fibre

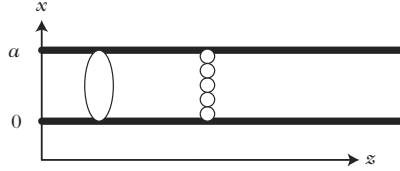
Before tackling the problem of the conical fibre, let us examine propagation in a metallic wave guide, using a simplified 1D model. To this end, we investigate the regions where the half-wavelength is less than (before cutoff), then greater than (after cutoff) the lateral dimensions of the wave guide [16].

Recall that for a transverse electric wave TE, guided by two conducting planes  $x = 0$  and  $x = a$ , the complex amplitude  $E(x)$  of the field  $E = E(x) \exp i(kz - \omega t)$  satisfies

$$\frac{d^2 E(x)}{dx^2} + \left( \frac{\omega^2}{c^2} - k^2 \right) E(x) = 0, \quad E(a) = E(0) = 0.$$

Propagating waves have the form (see Fig. 5.12)





**Fig. 5.12.** Electric field distribution in a metallic wave guide

$$E = C_1 \exp(ikx) + C_2 \exp(-ikx), \quad k^2 = \frac{\omega^2}{c^2} - k^2 \geq 0.$$

The boundary conditions require  $C_1 = -C_2$  and  $k = m\pi/a$ , where  $m$  is a nonzero integer. Consider the fundamental mode  $m = 1$ :

$$k^2 = \frac{\omega^2}{c^2} - \frac{\pi^2}{a^2}, \quad \text{or} \quad k = \frac{\omega}{c} \left(1 - \frac{\omega_C^2}{\omega^2}\right)^{1/2},$$

whence

$$k = \frac{\omega}{c} \left(1 - \frac{\lambda^2}{\lambda_C^2}\right)^{1/2} = \frac{2\pi}{\lambda} \left(1 - \frac{\lambda^2}{\lambda_C^2}\right)^{1/2},$$

where  $\omega_C = c\pi/a$  and  $\lambda_C = 2a$ . For  $\lambda > \lambda_C$ ,  $k$  is a complex number, and for  $\lambda \gg \lambda_C$ , it is pure imaginary, showing that there is amplitude attenuation during propagation.

### Application

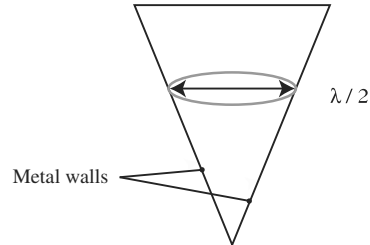
When we go from a wave guide with constant cross-section to a cone (see Fig. 5.13), we can calculate the attenuation of the wave by integrating over the cone. For example, for a cone of half-angle  $\alpha$  at the apex,

$$2\alpha = \frac{2}{10} \Rightarrow \text{amplitude attenuation} \sim 10^{-4.5} \Rightarrow \text{energy attenuation } 10^{-9}.$$

We thus find that a fine tip (i.e., with small  $\alpha$ ) capable of exploring correspondingly small structures will have a very low efficiency, whereas a less sharply tapered tip will be able to transmit more power. One is faced with a compromise between resolution and efficiency.

Energy that is not transmitted is dissipated by absorption (heating), whilst there is also a significant ‘reflection’ of the wave at the end of the fibre.

**Note.** We have neglected effects arising due to the skin thickness of the metal by assuming in the boundary conditions that the metal was perfect. Although this is not always true in the visible region of the spectrum, it is a reasonable hypothesis in the near IR.



**Fig. 5.13.** For a conical metal wave guide, the cutoff is reached when the width is less than  $\lambda/2$

### 5.4.3 Applications of Aperture SNOM

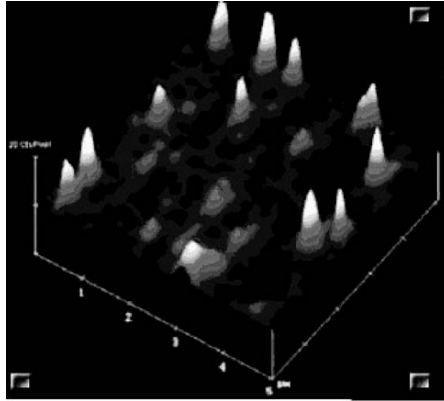
As mentioned above, aperture SNOM is the setup most commonly used today. In particular, commercially available instruments are based exclusively on this approach. The resolution of the latter instruments is around 50 nm, although the manufacturers guarantee ‘better than 100 nm’ and some research establishments can now achieve around 10 nm. They operate in the two modes, nanoscale light source and nanodetector, and they are always equipped with a feedback mechanism for the tip-sample separation which simultaneously confers the role of AFM upon them.

The simplest way to obtain recent illustrations of these setups is to visit the Websites of the main manufacturers, for the areas covered are vast, from the study of materials to electronic components, from soft matter (e.g., liquid crystals, polymers) to biology.

It is a unique advantage of this method that it supplies a highly localised (nanoscale) light source at the tip apex. This can be particularly useful in nanolithography, for example. Figure 5.14 provides a good illustration of the instrumental profile of such a microscope. The image obtained by a team at Harvard [11] shows single molecules scattered over a surface, which constitutes an excellent test of resolution.

With this example, we conclude the experimental part of the chapter. To complement the examples discussed above, the reader is encouraged to search the Web by putting the key words SNOM, apertureless SNOM, PSTM, etc., into an appropriate search engine. There are a great many examples and the field of applications is growing all the time.

Although we have attempted to quantify some of the phenomena covered above, the arguments have remained somewhat qualitative insofar as the description of the fields has been concerned. In the rest of this chapter, we shall go beyond these limitations by providing the tools required for a good understanding of the basic principles of near-field optics.



**Fig. 5.14.** Aperture SNOM image of single-molecule fluorescence. The field of view  $5\mu\text{m} \times 5\mu\text{m}$  allows an estimate of the resolution of this microscope ( $\sim 50\text{nm}$ ) [11]

## 5.5 Plane Wave Expansion. Diffraction Limit

In this second part of the chapter, we discuss the basic principles of near-field optics and, to use the recently-coined term, nano-optics. The techniques presented in the first part all rest upon the short-range interaction between a tip and an object, with a view to measuring the confined electromagnetic fields with a resolution that is not limited by diffraction. The aim of this theoretical part is threefold:

- To introduce the angular spectrum of plane waves in order to discuss both quantitatively and physically the origin of the diffraction limit, mentioned briefly in Sect. 5.1.1. We shall also discuss the role of evanescent waves and define the optical near field in terms of simple orders of magnitude.
- To use the elementary notions of electromagnetic radiation to introduce the concept of the optical near field from another point of view. We shall show that the laws of radiation also contain the diffraction limit, and we shall introduce the idea of the quasi-static (or electrostatic) limit, which can be used both to define the near field (without appealing to the idea of evanescent waves) and to set up simplified models, such as the one used in Sect. 5.3.2 to study the origin of contrast in an apertureless microscope.
- To discuss the problem of dipole emission by an atom or molecule in the vicinity of a nanostructure, using the classical model of the elastically bound electron. This simple approach is sufficient to introduce the ideas of modified radiative lifetime and frequency shift, as well as the question of radiative and non-radiative coupling. The near-field coupling between a single atom or molecule and a nanostructure is a recurring theme in nano-optics which we feel it important to introduce in this discussion.

The introduction given here can be complemented by the more complete discussion of the theory and modelling given in Chaps. 1–6 of [15].

In this section, we introduce the plane wave or angular spectrum expansion of a field propagating in vacuum. We use this to explain the diffraction limit and the notion of the optical near field in terms of the concept of evanescent waves.

### 5.5.1 Propagation of a Beam in Vacuum

We begin by establishing a general expression for the propagation in vacuum of a laterally confined monochromatic field, i.e., a beam. We choose the  $Oz$  axis in the direction of propagation and assume that the field is known in an arbitrary plane perpendicular to this axis, which can be designated  $z = 0$  without loss of generality. This scenario applies, for example, to a field that has crossed an object with known transparency, such as a diapositive, a diaphragm, or a slit like the one considered in Sect. 5.1.1. To simplify, we first treat the case of a scalar field in two dimensions. We then give the general result for a 3D vector field.

The monochromatic field with complex amplitude  $E(x, z)$  and frequency  $\omega$  [the time dependence  $\exp(-i\omega t)$  will be omitted throughout] obeys the Helmholtz equation

$$\nabla^2 E(x, z) + \frac{\omega^2}{c^2} E(x, z) = 0, \quad (5.4)$$

where  $c$  is the speed of light in vacuum. In the following, we put  $k = \omega/c = 2\pi/\lambda$ , where  $\lambda$  is the wavelength in vacuum. To obtain a well-posed problem, one must append two boundary conditions to this equation:

- we assume the field  $E(x, z = 0)$  is known,
- we assume that the field propagates towards  $z > 0$ .

We shall now show that it is possible to write down the general solution to this problem in the form of a superposition of plane waves, known as the plane wave or angular spectrum expansion.

In an arbitrary plane  $z > 0$ , we introduce the Fourier transform of the field with respect to the variable  $x$ , viz.,

$$E(x, z) = \int_{-\infty}^{+\infty} \tilde{E}(\alpha, z) \exp(i\alpha x) \frac{d\alpha}{2\pi}. \quad (5.5)$$

Substituting the expansion (5.5) into (5.4), we obtain the equation that must be satisfied by the Fourier transform of the field, viz.,

$$\frac{\partial^2 \tilde{E}(\alpha, z)}{\partial z^2} + \gamma^2 \tilde{E}(\alpha, z) = 0, \quad \text{where } \gamma^2 = k^2 - \alpha^2. \quad (5.6)$$

The general solution of this equation is

$$\tilde{E}(\alpha, z) = A(\alpha) \exp(i\gamma z) + B(\alpha) \exp(-i\gamma z), \quad (5.7)$$

where

$$\gamma = \begin{cases} \sqrt{k^2 - \alpha^2}, & \text{if } |\alpha| \leq k, \\ i\sqrt{\alpha^2 - k^2}, & \text{if } |\alpha| > k. \end{cases} \quad (5.8)$$

The boundary conditions now imply that  $B(\alpha) = 0$  and  $A(\alpha) = \tilde{E}(\alpha, z = 0)$ . We thus obtain an expression for the field in the form of an expansion in plane waves:

$$E(x, z > 0) = \int_{-\infty}^{+\infty} \tilde{E}(\alpha, z = 0) \exp[i\alpha x + i\gamma(\alpha)z] \frac{d\alpha}{2\pi}. \quad (5.9)$$

In this formula, we recall that the  $z$  component of the wave vector  $\gamma$  is the function of  $\alpha$  given by (5.8).

This result generalises in a straightforward manner to the case of a 3D monochromatic electromagnetic wave satisfying the vector Helmholtz equation

$$\nabla^2 \mathbf{E} + k^2 \mathbf{E} = 0.$$

We obtain, for all  $z > 0$ ,

$$\mathbf{E}(\mathbf{r}) = \int \tilde{\mathbf{E}}(\mathbf{K}, z = 0) \exp(i\mathbf{K} \cdot \mathbf{R} + i\gamma z) \frac{d^2 \mathbf{K}}{4\pi^2}, \quad (5.10)$$

where

$$\gamma(\mathbf{K}) = \begin{cases} \sqrt{k^2 - \mathbf{K}^2}, & \text{if } |\mathbf{K}| \leq k, \\ i\sqrt{\mathbf{K}^2 - k^2}, & \text{if } |\mathbf{K}| > k, \end{cases} \quad (5.11)$$

In this expression, we have used the notation  $\mathbf{r} = (x, y, z)$ ,  $\mathbf{R} = (x, y)$ , and  $\mathbf{K} = (\alpha, \beta)$  for the transverse wave vector. The range of integration is  $0 < |\mathbf{K}| < \infty$ .

Equation (5.10) represents the field as a linear superposition of plane waves with wave vector  $\mathbf{k} = (\mathbf{K}, \gamma)$  and amplitude equal to the Fourier transform  $\tilde{\mathbf{E}}(\mathbf{K}, z = 0)$  in the plane  $z = 0$ . The variable  $\mathbf{K}$  is then the spatial frequency associated with field variations in the plane  $z = 0$ . When the spatial frequency satisfies  $|\mathbf{K}| \leq k$  ( $k = \omega/c = 2\pi/\lambda$ ), the associated plane wave is propagative because the component  $\gamma$  of the wave vector in the  $z$  direction is real. For a high spatial frequency  $|\mathbf{K}| > k$ , this component  $\gamma$  of the wave vector is purely imaginary and the plane wave falls off exponentially in the  $z$  direction. We then have an evanescent wave. From this simple observation, one may deduce the following result: propagation in the vacuum behaves as a low-pass filter for spatial frequencies. We shall see that this filtering effect lies at the root of the resolution limit in classical optics.

### 5.5.2 Uncertainty Relations and Diffraction

Equation (5.10) describes propagation in vacuum, or any homogeneous medium, and involves no approximation. In particular, it contains diffraction effects, as we shall now show.

If for some reason the field  $\mathbf{E}(\mathbf{R}, z = 0)$  is laterally confined in a region of characteristic size  $\Delta R = \Delta x \Delta y$ , then the maximal spatial frequencies  $\alpha_{\max}$  and  $\beta_{\max}$  for which its spectrum takes significant values satisfy

$$\alpha_{\max} \Delta x \approx 2\pi, \quad \beta_{\max} \Delta y \approx 2\pi. \quad (5.12)$$

These relations, which we shall refer to as the uncertainty relations, are characteristic of the Fourier transform relationship between  $\mathbf{E}(\mathbf{R}, z = 0)$  and  $\tilde{\mathbf{E}}(\mathbf{K}, z = 0)$ .

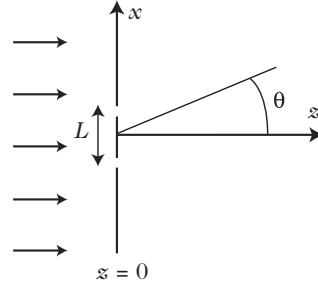
Let us assume for the moment that we observe the field far above the plane  $z = 0$ , in such a way that only the propagating waves contribute to the field. Consider a typical diffraction geometry. A slit of width  $L$  in the  $x$  direction and infinite in the  $y$  direction is cut into an opaque screen placed in the plane  $z = 0$ . If this screen is illuminated from the side  $z < 0$  by a monochromatic plane wave, the field in the plane  $z = 0$  is confined in the  $x$  direction with a characteristic scale  $L$ . The field propagating in the half-space  $z > 0$  is made up of plane waves with wave vectors  $(\alpha, \beta, \gamma)$ . The wave which deviates the furthest from the  $Oz$  axis has a wave vector in the  $x$  direction given by  $\alpha_{\max} \approx 2\pi/L$ , according to the uncertainty relation. (As the field is not confined in the  $y$  direction, only the spatial frequency  $\beta = 0$  contributes in this direction.) This simple argument shows that the transmitted field opens out in the  $(x, z)$  plane. If  $\theta$  is the angle defined by  $\alpha_{\max} = k \sin \theta$ , measuring the angular aperture of the beam, we obtain  $\theta \approx \lambda/L$ . We thus find, without calculation, the order of magnitude of the angular aperture of a beam due to diffraction.

### 5.5.3 Diffraction Limit

The above discussion also provides a qualitative understanding of how the resolution limit arises in classical optical systems, i.e., systems operating in the far field. If we are only concerned with the detection of propagating waves and we restrict to a 2D system in the  $(x, z)$  plane to simplify, the largest spatial frequency that can be detected is  $\alpha = k \sin \theta$ , where  $\theta$  is the collection angle of the detection system and  $\sin \theta$  is called the numerical aperture. According to the uncertainty relation, the smallest lateral variation in the field that can be detected is

$$\Delta x \approx \frac{2\pi}{\alpha} = \frac{\lambda}{\sin \theta}. \quad (5.13)$$

We thus obtain an order of magnitude estimate for the lateral resolution limit of classical imaging systems. For detection in a half-space ( $\sin \theta = 1$ ), this



**Fig. 5.15.** Typical geometry for diffraction due to two slits in an opaque screen. The system is invariant in the  $y$  direction

limit is of the same order as the wavelength of the light source. Its physical origin, like diffraction itself, lies in the uncertainty relation which relates the transverse width of a beam to its opening angle. It is thus commonly referred to as the diffraction limit.

The uncertainty relations (5.12) can be expressed in a quantum form by identifying the plane wave with wave vector  $\mathbf{k} = (\alpha, \beta, \gamma)$  with a photon with momentum  $\mathbf{p} = \hbar\mathbf{k}$ . In this case, the uncertainty relations can be written in the form

$$\Delta x \Delta p_x \approx h, \quad \Delta y \Delta p_y \approx h, \quad (5.14)$$

where  $\Delta p_x = \hbar\alpha_{\max}$  and  $\Delta p_y = \hbar\beta_{\max}$ . These are the Heisenberg uncertainty relations which arise naturally here as a consequence of a Fourier transform property. This form of the relations explains why, in some textbooks, the existence of the diffraction limit is presented as a consequence of the Heisenberg uncertainty relations. We shall see shortly that this point of view can lead to confusion when we speak of resolution beyond the diffraction limit.

The resolution limit can be obtained more precisely by calculating the diffraction pattern due to two parallel infinitely thin slits in an opaque screen, separated one from the other by a distance  $L$ , as shown in Fig. 5.15. Working in two dimensions with a scalar wave, the transmitted field in the plane  $z = 0^+$  can be written

$$E(x, z = 0) = A\delta(x - L/2) + A\delta(x + L/2), \quad (5.15)$$

where  $A$  is a constant and  $\delta$  the Dirac delta function. The associated angular spectrum is obtained directly as

$$\tilde{E}(\alpha, z = 0) = A \exp(i\alpha L/2) + A \exp(-i\alpha L/2). \quad (5.16)$$

Suppose now that an ideal optical system produces an image in an arbitrary image plane of the field in the plane  $z = 0$ , collecting only the propagating waves with a given numerical aperture  $\sin \theta$ . The field in the image plane is then

$$\begin{aligned}
E_{\text{det}}(x) &= \int_{-k \sin \theta}^{+k \sin \theta} \tilde{E}(\alpha, z=0) \exp(i\alpha x) \frac{d\alpha}{2\pi} \\
&= 2A \frac{\sin \theta}{\lambda} \left\{ \text{sinc} \left[ k \sin \theta (x + L/2) \right] + \text{sinc} \left[ k \sin \theta (x - L/2) \right] \right\},
\end{aligned} \tag{5.17}$$

where  $\text{sinc}(x) = \sin x/x$  is the cardinal sine function. The field in the detection plane is thus composed of two cardinal sine functions centered at  $x = -L/2$  and  $x = +L/2$ , respectively. When we measure the field, what is the condition for being able to separate the two slits in the object plane? To answer this question, we may apply the Rayleigh criterion, which says that the two slits can be separated if the maximum of one of the diffraction patterns coincides with the minimum of the other. In the case where we have two sinc functions, we may say that the first zero of one of the functions must coincide with the maximum of the other. This happens when  $kL \sin \theta = \pi$ , whence

$$L = \frac{\lambda}{2 \sin \theta}. \tag{5.18}$$

According to the Rayleigh criterion,  $L$  is the smallest separation for which one may distinguish the two slits. It is therefore the resolution limit for classical systems. For a numerical aperture  $\sin \theta = 1$ , we obtain  $L = \lambda/2$ , which is around 200 nm when using a visible wavelength.

## 5.6 Beyond the Diffraction Limit: Near Field and Evanescent Waves

### 5.6.1 Evanescent Waves. Length Scales

The analysis in the last section showed that the high spatial frequencies of the field, corresponding to subwavelength lateral spatial variations in the field  $\mathbf{E}(\mathbf{R}, z=0)$ , are exponentially attenuated as one moves away from the plane  $z=0$ . If we wish to exploit the rapid variations of the field, we must therefore move in closer and detect the evanescent waves. In that case, we may expect to measure spatial variations of the field on length scales much shorter than  $\lambda$ . This is one of the challenges of near-field optical microscopy.

Let us suppose that the field in the plane  $z=0$  varies laterally on a length scale  $\Delta x \ll \lambda$ . At what distance must we detect the field in order to observe these variations? The associated spatial frequencies are of the order of  $\alpha = 2\pi/\Delta x$ . The associated plane waves fall off exponentially according to

$$\exp\left(-z\sqrt{\alpha^2 - k^2}\right) \approx \exp(-|\alpha|z) = \exp(-z/\delta), \tag{5.19}$$

where  $\delta = \Delta x/2\pi$ . For concreteness, imagine that we wish to measure a lateral variation  $\Delta x = 50$  nm. In this case, we must approach to a distance of the order of  $z \approx 10$  nm.



This simple argument can be used to make a first definition of the near field: the near field is the region where evanescent waves contribute significantly to the electromagnetic field. This region has a spatial extent of  $\delta \approx \Delta x/2\pi$ , where  $\Delta x$  is the characteristic length scale of the lateral variations of the field.

It is important to note that the characteristic length scale  $\delta$  of the near field is independent of the wavelength  $\lambda$ . On very short length scales compared with  $\lambda$ , the wavelength is no longer a relevant quantity. We shall come back to this point in the next section when we discuss the quasi-static limit.

### 5.6.2 Uncertainty Relations Revisited

It may seem paradoxical to have shown that the uncertainty relations (5.12) explained in a certain sense the diffraction limit of classical optics, and then demonstrated from the same starting point, viz., the plane wave expansion, that it is possible to go beyond this limit. In fact, there is no contradiction here, as explained in detail in [22]. If we rewrite the conditions on the transverse wave vector in the form

$$\alpha_{\max}\Delta x \approx 2\pi, \quad \beta_{\max}\Delta y \approx 2\pi, \quad (5.20)$$

$$\alpha^2 + \beta^2 + \gamma^2 = k^2, \quad (5.21)$$

we see that, by allowing  $\gamma^2$  to be negative, which amounts to admitting the detection of evanescent waves, the spatial frequencies  $\alpha$  and  $\beta$  can be arbitrarily large whilst still satisfying the dispersion relation (5.21). Arbitrarily large spatial frequencies correspond to arbitrarily small length scales  $\Delta x$  and  $\Delta y$ , according to the uncertainty relations (5.20). In particular, we see that there is a risk of confusion in presenting the diffraction limit as a consequence of the Heisenberg uncertainty relations (5.14), which are none other than the relations (5.20) rewritten in a different form. Indeed, as we have just seen, it is possible to obtain a resolution beyond the diffraction limit by measuring evanescent waves, whilst still satisfying both the Heisenberg relations (5.14) and the dispersion relation (5.21).

## 5.7 Electromagnetic Radiation. Near Field and Far Field

### 5.7.1 Radiation from an Elementary Source (Electric Dipole)

The notion of optical near field can be introduced from a different standpoint to the one we have used up to now, without bringing in the idea of evanescent waves. To this end, we turn to the fundamental features of electromagnetic radiation. The simplest approach consists in considering an elementary source, namely, an electric dipole. Such a source is interesting for two main reasons:

- most sources can be treated as electric dipoles when they are much smaller than the wavelength,
- the radiation of an electric dipole is a basic element in understanding emissions from a single atom or molecule.

Concerning the second point, the ability to measure the emission of single atoms and molecules constitutes one of the main advances of nano-optics, as we shall see at the end of the chapter.

Consider a monochromatic dipole of frequency  $\omega$ , placed at the origin of the coordinate system. This dipole can be imagined either as a material particle of negligible size, in particular, very small compared with the light wavelength, in which oscillating currents are excited, or simply as an oscillating point charge. In the second view, the associated dipole moment is then  $\mathbf{p} = q\mathbf{R}$ , where  $q$  is the charge and  $\mathbf{R}$  its position. The electric field radiated to point  $\mathbf{r}$  is then [8, 14]

$$\mathbf{E}(\mathbf{r}) = \frac{k^2}{4\pi\epsilon_0} \frac{\exp(ikr)}{r} \left\{ \mathbf{p} - (\mathbf{p} \cdot \mathbf{u})\mathbf{u} - \left( \frac{1}{ikr} + \frac{1}{k^2r^2} \right) [\mathbf{p} - 3(\mathbf{p} \cdot \mathbf{u})\mathbf{u}] \right\}, \quad (5.22)$$

where  $\mathbf{u} = \mathbf{r}/r$  is a unit vector in the direction of observation.

Two comments are in order concerning this expression for the dipole field:

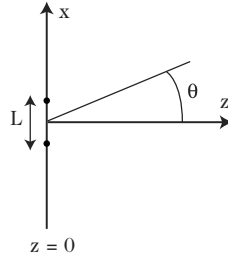
- There are three terms with different spatial dependences, proportional to  $r^{-1}$ ,  $r^{-2}$ , and  $r^{-3}$ . The first term is the far-field term, the only one contributing to the energy flux radiated at great distances. The last term is the near-field term, dominating at short distances. We shall discuss their properties below. Note that there is also an intermediate term, which shows that the near field in the usual sense of nano-optics and near-field optics cannot be identified with the contribution complementary to the far field.
- The spatial structure of the field is highly anisotropic, especially in the near field. This anisotropy of the dipole field underlies many so-called polarisation effects encountered in nano-optics.

### 5.7.2 Far-Field Radiation. Diffraction Limit Revisited

We shall show that, by restricting measurement to radiation in the far field, we retrieve the resolution limit of classical optics given by (5.18). Consider two identical monochromatic point sources (dipoles with dipole moment  $\mathbf{p}$ ) placed in the plane  $z = 0$  and separated by a distance  $L$ . We assume that the dipoles are located on the  $Ox$  axis and oriented in the  $y$  direction, working entirely in the  $(Ox, Oz)$  plane to simplify, as shown in Fig. 5.16.

The electric field radiated into the far field in the direction  $\theta$  can then be written

$$\mathbf{E}(\mathbf{r}) = \frac{\mu_0\omega^2}{2\pi} \frac{\exp(ikr)}{r} \cos\left(k \sin\theta \frac{L}{2}\right) \mathbf{p}. \quad (5.23)$$



**Fig. 5.16.** Far-field radiation from two point sources. We seek the field radiated into the far field in the plane  $(Ox, Oz)$

By examining the resulting field (or its intensity) in the far field, between angles  $-\theta_{\max}$  and  $+\theta_{\max}$ , we observe an interference pattern: as  $\theta$  varies, the intensity is modulated as  $\cos^2(k \sin \theta L/2)$ . A simple resolution criterion is then possible: under what conditions can the two sources no longer be distinguished one from the other? Now this happens when we can no longer measure any fringe in the interference pattern. So the limit is obtained when there is exactly one fringe between  $-\theta_{\max}$  and  $+\theta_{\max}$ . This condition can be written

$$k \sin \theta_{\max} \frac{L}{2} = \frac{\pi}{2}, \quad \text{i.e., } L = \frac{\lambda}{2 \sin \theta_{\max}}. \quad (5.24)$$

This relation gives the minimum value of the separation  $L$  between the two sources for which a far-field measurement is able to distinguish them. We have thus obtained the resolution limit of the system and retrieved the result (5.18) without appealing to the notions of evanescent waves and propagating waves and without mentioning diffraction.

The main conclusion from this simple calculation is that the resolution limit in classical optics originates in the fact that measurements are made in the far field. We note that far-field conditions can be obtained, for example, in the focal plane of a convergent lens. The existence of this limit is perfectly described by the basic laws of electromagnetic radiation.

### 5.7.3 Near-Field Radiation. Quasi-Static Limit

We now consider the case where the observation distance  $r$  is very small compared with the wavelength  $\lambda$ . The electric field is obtained from (5.22) in the limit  $kr \rightarrow 0$ , whence

$$\mathbf{E}_{\text{qs}}(\mathbf{r}) = \frac{3(\mathbf{p} \cdot \mathbf{u})\mathbf{u} - \mathbf{p}}{4\pi\epsilon_0 r^3}. \quad (5.25)$$

Physically, we are concerned with the situation in which the frequency  $\omega$  is fixed whilst the distance  $r$  tends to 0. Mathematically, the limit  $kr \rightarrow 0$  can also be interpreted as the case where the speed of light  $c$  tends to infinity

whilst  $r$  and  $\omega$  remain fixed. This is the quasi-static limit, in which we neglect retardation effects. Although the frequency  $\omega$  is fixed, and corresponds in our situation to optical and near IR frequencies, we then retrieve exactly the same equations as in electrostatics. In particular, (5.25) is the same as the equation for the electric dipole field obtained in electrostatics, although in the present formula  $\mathbf{E}_{\text{qs}}$  and  $\mathbf{p}$  both oscillate at frequency  $\omega$ , which is of the order of  $10^{14}$  to  $10^{15}$  Hz in visible and near-IR optics!

We can now give a second definition of the optical near field: the near field is the region where quasi-static contributions dominate. For point (dipole) sources, this region is the one where the field is dominated by terms going as  $r^{-3}$ .

This second definition of the near field provides a point of view that differs, and indeed is complementary to the one used in Sect. 5.6.1. In particular, it shows that, in the context of the quasi-static limit, the wavelength  $\lambda$  is no longer a relevant parameter. The important quantities are the frequency  $\omega$ , which determines the optical response of the observed objects, and the observation distance. This second point of view thus explains why the wavelength does not appear in (5.19), which defined the extent of the near field from the standpoint of evanescent waves.

#### 5.7.4 Towards a Model

The basic concepts introduced above help us to understand the origin of the fundamental limits defining classical optics, and at the same time, the operating principles underlying the techniques of near-field optics presented earlier. They provide a qualitative approach to problems and order of magnitude estimates. In a second stage, one can begin to model the techniques of near-field optics in a more detailed way, aiming at a quantitative description of image formation processes, i.e., the origin of contrast, and the possibility of optimising experimental setups. This is not the place to enter into a full discussion of more advanced methods. Historically, the first models were developed for the PSTM [21] and the aperture SNOM [23], using the tools presented in this chapter, i.e., angular spectrum and dipole radiation. The reader will find a review of the relevant physical models for apertureless SNOM in [3]. More complete discussions containing a review of basic ideas, together with their use in the appropriate models, and the role played by numerical simulation in the area of near-field optics can be found in [9, 10].

### 5.8 Dipole Emission Near a Nanostructure

In this last section, we discuss the problem of emission from an atom or molecule<sup>3</sup> in the vicinity of a nanostructure. We shall show that a simple classical

<sup>3</sup> Throughout this section, we shall use the word ‘atom’ to denote either an atom or a molecule.

model suffices to understand how the radiative lifetime and emission frequency are expected to evolve, and we shall also briefly discuss the connection with the quantum approach. We tackle the question of competition between radiative and non-radiative (absorptive) coupling. For a more complete and fully pedagogical approach to these matters, the reader is recommended to consult [17]. A higher level discussion referring to the latest results in this field can be found in the review paper [14].

To begin with, we present the classical model of the elastically bound electron and introduce the idea of radiative damping. We shall show that radiative damping is responsible for the finite lifetime of the dipole emission and also for the appearance of a frequency shift. In a second stage, we show how the lifetime and frequency shift are modified by interaction with a nanostructure located in the vicinity of the emitting atom. Finally, we examine two simple examples which illustrate the method and the main physical effects. In particular, we introduce the idea of radiative and non-radiative coupling.

### 5.8.1 Radiative Damping of Dipole Emission

We consider an electron in the outer shell of an atom and model the way it is bound to the atom by an elastic restoring force  $\mathbf{F}_b(t) = -m\omega_0^2\mathbf{R}(t)$ , where  $m$  is the electron mass,  $\mathbf{R}$  the deviation from the equilibrium trajectory, and  $\omega_0$  the resonance frequency of the bond.

To begin with, we assume that, after excitation, the electron oscillates freely about its equilibrium trajectory at a frequency  $\omega_0$ . This electron in accelerated motion must radiate, and hence lose energy. Let us calculate the power lost due to radiation and calculate the temporal evolution of the energy of this oscillator.

The dipole moment due to the moving electron is  $\mathbf{p}(t) = -e\mathbf{R}(t)$ . To simplify notation, we write  $\mathbf{p}(t) = \mathbf{p}\exp(-i\omega_0 t)$  and omit the factor  $\exp(-i\omega_0 t)$  in the following, as we shall do for all time-dependent quantities. The time average of the power radiated by the dipole is

$$P = \frac{\mu_0\omega_0^4}{12\pi c} |\mathbf{p}|^2. \quad (5.26)$$

This power is obtained from the expression for the radiated field in (5.22), keeping only the far-field term. It can be used to calculate the time-averaged value of the Poynting vector  $\mathbf{II}(\mathbf{r}) = \varepsilon_0 c^2 |\mathbf{E}(\mathbf{r})|^2 / 2$ , whose flux out through a sphere of radius  $r$  gives the radiated power  $P$ .

The total energy contained in the dipole, i.e., the sum of the kinetic energy and potential energy, which have the same time-averaged value for a harmonic oscillator, is

$$U = U_{\text{kin}} + U_{\text{pot}} = 2U_{\text{kin}} = \frac{m\omega_0^2}{2} |\mathbf{R}|^2. \quad (5.27)$$

Comparing the expressions (5.26) and (5.27), we see that we have  $P = \Gamma_0 U$ , where

$$\Gamma_0 = \frac{e^2 \omega_0^2}{6\pi m \varepsilon_0 c^3}. \quad (5.28)$$

The factor  $\Gamma_0$  is called the free-space damping rate.

The dipole loses energy by radiation. Assuming that the decrease in energy  $U$  is slow compared with  $2\pi/\omega_0$ , we may write

$$\frac{dU}{dt} = -P = -\Gamma_0 U.$$

This implies that the dipole energy decreases exponentially according to a law of the form

$$U(t) = U_0 \exp(-\Gamma_0 t).$$

The characteristic dipole emission time  $\tau_0 = 1/\Gamma_0$  is known as the free-space radiative lifetime. This quantity is the classical analogue of the radiative lifetime of an excited level of an atom in quantum physics.

### 5.8.2 Free-Space Dipole Emission

Given that the oscillating electron loses energy by radiation, we may seek the equation of motion of the dipole moment  $\mathbf{p}$  and solve it. We then obtain the expression for the radiative lifetime and show that the oscillation frequency is slightly shifted away from  $\omega_0$  due to the damping.

Applying Newton's second law to the electron and taking into account the restoring force and damping term, we find that

$$\frac{d^2 \mathbf{p}}{dt^2} + \Gamma_0 \frac{d\mathbf{p}}{dt} + \omega_0^2 \mathbf{p} = 0. \quad (5.29)$$

We seek a solution of the form  $\mathbf{p}(t) = \mathbf{p} \exp(-i\Omega t)$ , where  $\Omega$  may be complex. This leads to a possible solution for

$$\Omega = \frac{1}{2} \left( \sqrt{4\omega_0^2 - \Gamma_0^2} - i\Gamma_0 \right),$$

with the proviso that the fall-off is slow, i.e.,  $\Gamma_0 < 2\omega_0$ , which we shall check later in a specific example. We then obtain

$$\mathbf{p} = \mathbf{p} \exp(-\Gamma_0 t) \exp(-i\omega t), \quad (5.30)$$

which gives once again the free-space radiative lifetime as  $\tau_0 = 1/\Gamma_0$  and an oscillation frequency

$$\omega = \omega_0 \sqrt{1 - \frac{\Gamma_0^2}{4\omega_0^2}} . \quad (5.31)$$

The oscillation frequency is thus shifted slightly from the resonance frequency  $\omega_0$  of the bond. This shift is a direct consequence of radiative damping.

It is useful at this stage to calculate some orders of magnitude. For emission in the visible, we have  $\omega_0 \sim 10^{15}$  Hz. Equation (5.28) gives  $\Gamma_0 \sim 10^9$  Hz numerically, and hence a radiative lifetime of  $\tau_0 \sim 10^{-9}$  s. The slow damping condition  $\Gamma_0 < 2\omega_0$  is indeed satisfied. Moreover the frequency shift given by (5.31) is extremely small and the approximation  $\omega \approx \omega_0$  is generally justified in practice.

### 5.8.3 Dipole Emission Near an Object

Let us now investigate how the dynamics of the oscillating dipole is affected by the presence of a nearby object, e.g., a surface, a small particle, or a microscope tip. The physical origin of such an effect is the interaction of the electron with part of its own radiation which has been ‘reflected’ by the object. We shall show that the radiative lifetime and emission frequency are both modified with respect to their free-space values.

Let  $\mathbf{E}_{\text{loc}}$  be the electric field at the position of the electron resulting from ‘reflection’ by the nearby object, sometimes called the local field. The equation of motion of the dipole moment  $\mathbf{p}$  is changed by the presence of a new force, becoming

$$\frac{d^2\mathbf{p}}{dt^2} + \Gamma_0 \frac{d\mathbf{p}}{dt} + \omega_0^2 \mathbf{p} = \frac{e^2}{m} \mathbf{E}_{\text{loc}} . \quad (5.32)$$

Note that the magnetic force is negligible, except for a relativistic electron (which is not the case here). This explains why only the electric force is included in this equation. As before, we seek a solution of the form  $\mathbf{p}(t) = \mathbf{p} \exp(-i\Omega t)$ , whereupon we obtain

$$\mathbf{p} = \mathbf{p} \exp(-\Gamma t) \exp(-i\omega t) . \quad (5.33)$$

With the weak damping condition  $\Gamma_0 < 2\omega_0$ , the damping rate is given by

$$\Gamma = \Gamma_0 + \frac{e^2}{m\omega_0|\mathbf{p}|^2} \text{Im}(\mathbf{p}^* \cdot \mathbf{E}_{\text{loc}}) , \quad (5.34)$$

and the frequency shift by

$$\Delta\omega = \omega - \omega_0 = -\frac{\Gamma_0^2}{8\omega_0} - \frac{e^2}{2m\omega_0|\mathbf{p}|^2} \text{Re}(\mathbf{p}^* \cdot \mathbf{E}_{\text{loc}}) , \quad (5.35)$$

where Re and Im denote the real and imaginary parts of the following quantities, respectively, and the asterisk denotes the complex conjugate.

The damping rate and frequency shift obtained here differ from those found in free space by a quantity that is directly related to the local field, and hence to the influence the dipole has on itself via the nearby object. In fact, these changes in the damping rate (and hence the radiative lifetime) and the frequency shift are independent of the dipole itself, as we shall show.

When the object is illuminated by the dipole (placed at point  $\mathbf{r}_0$ ), the resulting field at an arbitrary point  $\mathbf{r}$  is linearly related to the dipole source, by virtue of the linearity of Maxwell's equations. This relationship is usually expressed in the form of a tensor relation

$$\mathbf{E}_{\text{loc}}(\mathbf{r}) = \mathbf{S}(\mathbf{r}, \mathbf{r}_0, \omega_0) \mathbf{p}, \quad (5.36)$$

where the tensor  $\mathbf{S}$  is called Green's tensor or the linear susceptibility of the field. The fact that the relation is tensorial simply expresses the fact that the field and the dipole moment are not generally collinear. Replacing  $\mathbf{E}_{\text{loc}}$  in (5.34) and (5.35) by its expression in terms of  $\mathbf{S}$ , we obtain

$$\frac{\Gamma}{\Gamma_0} = 1 + \frac{6\pi\epsilon_0 c^3}{\omega_0^3} \text{Im} \left[ \mathbf{u} \cdot \mathbf{S}(\mathbf{r}_0, \mathbf{r}_0, \omega_0) \mathbf{u} \right], \quad (5.37)$$

$$\frac{\Delta\omega}{\Gamma_0} = -\frac{\Gamma_0}{8\omega_0} - \frac{3\pi\epsilon_0 c^3}{\omega_0^3} \text{Re} \left[ \mathbf{u} \cdot \mathbf{S}(\mathbf{r}_0, \mathbf{r}_0, \omega_0) \mathbf{u} \right], \quad (5.38)$$

where  $\mathbf{u}$  is the unit vector in the direction of the dipole, i.e.,  $\mathbf{p} = p\mathbf{u}$ . These two relations yield the modifications in the emission rate and frequency shift (as a fraction of their free-space values) induced by interaction with the object. These changes only depend on Green's tensor  $\mathbf{S}(\mathbf{r}_0, \mathbf{r}_0, \omega_0)$ , which expresses the electrodynamic response of the object<sup>4</sup> when it is illuminated by a dipole placed at  $\mathbf{r}_0$ .

#### 5.8.4 Link with the Quantum Approach

Equations (5.37) and (5.38) were obtained from a classical model. This raises the question of the validity of these results, given that radiative emissions (spontaneous emission) from an atom are generally tackled using quantum theory.

In the framework of quantum physics, the equivalent of the classical emission rate that we have introduced above is the spontaneous emission rate. For a two-level atom with ground state  $|a\rangle$  and excited state  $|b\rangle$ , the spontaneous emission rate due to coupling with a monochromatic electromagnetic field of frequency  $\omega_0$  is calculated using perturbation theory and Fermi's golden rule. A summary of this calculation can be found in [17]. For an atom in free space (vacuum), we obtain

<sup>4</sup> More precisely,  $\mathbf{S}$  is the modification of the empty-space Green tensor due to the presence of the object.



$$\Gamma_0^{b \rightarrow a} = \frac{4\omega_0^3}{3\hbar c^3} |\langle a | \mathbf{p} | b \rangle|^2, \quad (5.39)$$

where  $\langle a | \mathbf{p} | b \rangle$  is the matrix element of the electric dipole moment operator between states  $|a\rangle$  and  $|b\rangle$ . Note that this differs from the classical result (5.28), in particular, by the presence of Planck's constant. The spontaneous emission rate is a purely quantum quantity here. However, when we calculate the modification in the emission rate due to the presence of an arbitrary object characterised by Green's tensor  $\mathbf{S}$ , we obtain the same expression (5.37) for the normalised emission rate as we did from the classical calculation. (The calculation uses Fermi's golden rule once again. More details can be found in [17].) We thus find

$$\left. \frac{\Gamma^{b \rightarrow a}}{\Gamma_0^{b \rightarrow a}} \right|_{\text{quantum}} = \left. \frac{\Gamma}{\Gamma_0} \right|_{\text{classical}}. \quad (5.40)$$

We may conclude that, as far as the normalised emission rate is concerned, the classical model provides the correct result.

With regard to the frequency shift, (5.38) describes the so-called classical shift. In the framework of quantum physics, other shifts arise from the coupling between the atom and electromagnetic fluctuations of the vacuum (see [17] for a pedagogical account). These frequency shifts of quantum origins may be greater than the classical shift. The investigation of the frequency shift given in the classical framework is thus inadequate. We therefore restrict discussion to the modification of the emission rate in the following.

### 5.8.5 A Simple Example: Dipole Emission Near a Plane Mirror

The modification in the emission rate of a molecule in the vicinity of a structure was first demonstrated in the 1970s, in particular, through experiments by Drexhage (1970). These experiments were then explained, using the classical theory described above, by Chance, Prock and Silbey [4]. The object here is a plane metal mirror. Concerning the radiative lifetime, two effects were demonstrated experimentally and described by the model:

- Far from the plane (in the far field), the normalised lifetime oscillates with period  $\lambda_0/2$ , where  $\lambda_0 = 2\pi c/\omega_0$  is the emission wavelength.
- At short distances (in the near field), the lifetime falls off and goes to zero at contact.

We shall show how the classical model can be used to describe this behaviour, at least qualitatively.

Consider an atom placed in front of a perfectly conducting metallic plane (a mirror) denoted by  $z = 0$ . We shall calculate the modification in the dipole emission rate (or the radiative lifetime) induced by the presence of this mirror. The simplest case corresponds to a dipole oriented parallel to the plane  $z = 0$  and located at a distance  $d$  which we suppose first to be large compared with

the emission wavelength. The mirror has the same effect as an image dipole of dipole moment  $-\mathbf{p}$  placed at  $z = -d$ . Indeed, this guarantees the boundary condition whereby the electric field must vanish on the mirror surface. The local field radiated by the image dipole is then

$$\mathbf{E}_{\text{loc}}(z = d) = -\frac{\mu_0 \omega_0^2}{4\pi} \frac{\exp(2ikd)}{2d} \mathbf{p}, \quad k = \frac{\omega_0}{c} = \frac{2\pi}{\lambda_0}. \quad (5.41)$$

Substituting this expression into (5.34), we obtain the modification of the emission rate directly as

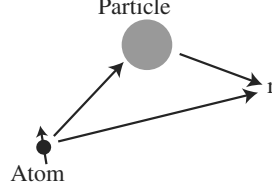
$$\frac{\Gamma}{\Gamma_0} = 1 - \frac{3}{2} \text{sinc}(2kd), \quad (5.42)$$

where  $\text{sinc}(x) = \sin x/x$  is the cardinal sine function. This expression is only valid when  $d \gg \lambda_0$ . It shows that the normalised emission rate (and hence the lifetime) oscillates as  $d$  varies. The oscillations have period  $\lambda_0/2$  and vanish as  $d \rightarrow \infty$ , whereupon we retrieve the free-space emission rate. Physically, these oscillations originate in interference between the field emitted by the dipole and the field reflected by the mirror. Note in particular that the emission rate can be greater or less than its free-space value, depending on the distance  $d$ . This result explains qualitatively the experimentally observed behaviour at large distances [4].

At shorter distances (in the near-field region of the mirror), the experiment shows that the lifetime tends to zero. The interaction of the emitting dipole with the mirror can still be replaced by the interaction between this dipole and its image located at  $z = -d$ . However, in this case, we retain only the near-field (quasi-static) term in the reflected field. The method then closely resembles the one discussed in Sect. 5.3.2. The short-distance behaviour is recovered qualitatively and we shall not give the details of this calculation in the case of the plane mirror (but see [4, 17]). However, the more general problem of the interaction between an emitting dipole and another dipole in the near field will be the subject of the next section. The case of the plane mirror, in which the second dipole is the image of the first, can then be deduced as a special case.

### 5.8.6 Dipole Emission Near a Nanoparticle. Radiative and Non-Radiative Coupling

The interaction between an atom and a metallic or dielectric nanoparticle lies at the heart of many applications of nano-optics, e.g., single-molecule spectroscopy by local probe microscopy, subwavelength guiding of light by metal nanoparticles, recognition of adsorbed molecules on resonant dielectric particles, etc. In order to obtain a first glimpse of the modifications to the radiative behaviour of an atom in close proximity to a nanoparticle, it is useful to consider the simple case of a particle which is itself behaving as an



**Fig. 5.17.** Dipole emission in the vicinity of a nanoparticle. The radiation at the point  $\mathbf{r}$  contains two contributions: one direct, the other passing via the particle

electric dipole (induced by the external field). This occurs when the radius  $a$  of the particle satisfies  $a \ll \lambda_0$  and  $a < l$ , where  $l$  is the distance between the atom and the surface of the particle.

In this section, we calculate the modification in the emission rate due to the presence of the particle. We also discuss the effects of scattering and absorption by the particle of the field emitted by the atom, i.e., radiative and non-radiative coupling, respectively.

Suppose the atom is placed at  $\mathbf{r}_0$  and the particle is centered at  $\mathbf{r}_p$ . As we have seen, the emission rate calculation depends on knowing Green's tensor in the presence of the particle [see (5.37)]. The field radiated by a point dipole placed at  $\mathbf{r}_0$  contains two contributions: direct radiation as in empty space, and radiation through mediation by the particle, which is polarised and then itself radiates (see Fig. 5.17).

Denoting the empty-space Green tensor by  $\mathbf{G}_0$ , the total Green tensor is

$$\mathbf{G}(\mathbf{r}, \mathbf{r}_0, \omega_0) = \mathbf{G}_0(\mathbf{r}, \mathbf{r}_0, \omega_0) + \mathbf{G}_0(\mathbf{r}, \mathbf{r}_p, \omega_0) \alpha(\omega_0) \varepsilon_0 \mathbf{G}_0(\mathbf{r}_p, \mathbf{r}_0, \omega_0), \quad (5.43)$$

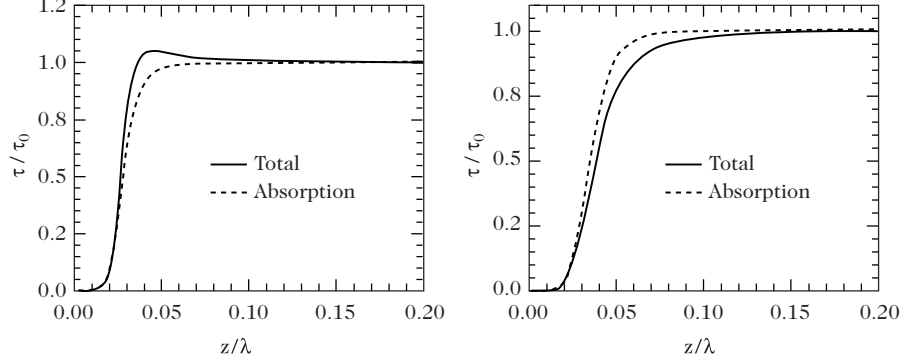
where  $\alpha(\omega)$  is the polarisability of the particle, so that the dipole moment induced by an external field  $\mathbf{E}_{\text{ext}}(\omega)$  is given by

$$\mathbf{p}_{\text{ind}}(\omega) = \alpha(\omega) \varepsilon_0 \mathbf{E}_{\text{ext}}(\omega).$$

For a particle of radius  $a$  and dielectric constant  $\varepsilon(\omega)$ , the polarisation is given by [8]

$$\alpha(\omega) = \frac{\alpha_0(\omega)}{1 - ik^3/6\pi \alpha_0(\omega)}, \quad \alpha_0(\omega) = 4\pi a^3 \frac{\varepsilon(\omega) - 1}{\varepsilon(\omega) + 2}. \quad (5.44)$$

In this expression,  $\alpha_0$  is the usual Clausius–Mossotti polarisability. The expression given here has been adjusted to remain consistent with the fact that, for a non-absorbing particle, i.e., with real  $\varepsilon(\omega)$ , the polarisability must nevertheless include a nonzero imaginary part which reflects energy losses due to scattering. (A plane wave incident on the particle loses energy which is reradiated by the particle in other directions [8].) In technical terms, one says that the polarisability must be consistent with the optical theorem which expresses this point mathematically [1]. For a correct treatment of the emission



**Fig. 5.18.** Normalised lifetime of an atom in close proximity to a silver nanoparticle of radius  $a = 10$  nm as a function of the distance  $z$  to the centre of the particle. The emission wavelength is  $\lambda = 612$  nm. *Left:* Dipole moment of the atom oriented perpendicularly to the atom–particle direction. *Right:* Dipole moment of the atom pointing towards the particle. *Dashed curve:* Modification of the lifetime due only to absorption. At short distances, absorption predominates

rate problem, this correction must be taken into account at the outset, and this is why we have mentioned it. However, this rather technical point is not a fundamental issue in the present approach. Moreover, for an absorbing particle, e.g., a metal particle, the relation  $\alpha \approx \alpha_0$  is in practice a very good approximation.

The empty-space Green tensor follows directly from the expression (5.22) for the field radiated into the vacuum by the dipole. We obtain

$$\mathbf{G}_0(\mathbf{r}, \mathbf{r}', \omega_0) = \frac{k^2}{4\pi\epsilon_0} \frac{\exp(ikR)}{R} \left[ \mathbf{I} - \mathbf{u}\mathbf{u} - \left( \frac{1}{ikR} + \frac{1}{k^2R^2} \right) (\mathbf{I} - 3\mathbf{u}\mathbf{u}) \right], \quad (5.45)$$

where the tensor  $\mathbf{u}\mathbf{u}$  is such that  $(\mathbf{u}\mathbf{u})\mathbf{p} = (\mathbf{p} \cdot \mathbf{u})\mathbf{u}$  and  $R = |\mathbf{r} - \mathbf{r}'|$ . Finally, the expression for the tensor  $\mathbf{S} = \mathbf{G} - \mathbf{G}_0$ , which is the modification of the Green tensor induced by the presence of the particle, follows from (5.43) and (5.45):

$$\mathbf{S}(\mathbf{r}_0, \mathbf{r}_0, \omega_0) = \alpha(\omega_0)\epsilon_0 \mathbf{G}_0(\mathbf{r}_0, \mathbf{r}_p, \omega_0) \mathbf{G}_0(\mathbf{r}_p, \mathbf{r}_0, \omega_0). \quad (5.46)$$

Substituting this expression into (5.37), we obtain the normalised emission rate, and hence the modification in the lifetime. An example is shown in Fig. 5.18. The atom emits at wavelength  $\lambda_0 = 612$  nm and is situated at a distance  $z$  from a silver nanoparticle of radius  $a = 10$  nm. The graphs show the normalised lifetime  $\tau/\tau_0 = (\Gamma/\Gamma_0)^{-1}$  as a function of the distance  $z$  for two orientations of the dipole moment  $\mathbf{p}$  of the atom.

We observe that, at short distances, the lifetime falls off sharply and ends up going to zero at contact. This reduction in the lifetime, which corresponds to an increase in the emission rate, can be attributed to two physical causes:

- When the atom emits alone in empty space, part of the emitted field is non-propagative and remains confined in the vicinity of the atom. [This field corresponds to the terms going as  $r^{-2}$  and  $r^{-3}$  in the dipole field of (5.22).] When it emits near the particle, part of this non-propagating field is converted to a propagating field by coupling with the particle. The atom thus loses more power by radiation into the far field and its radiative lifetime is reduced. One then speaks of radiative coupling.
- Part of the radiation emitted by the atom is absorbed by the particle when it is made from some material absorbing at the emission wavelength  $\lambda_0$ . The presence of absorption creates a new way for the atom to lose power, and this also tends to diminish its radiative lifetime. In this case, the lost energy is not radiated into the far field, but is transformed into heat energy in the particle. One then speaks of non-radiative coupling.

In the model used here, it is easy to carry out separate calculations of the modifications to the emission rate due to the two types of coupling. In Fig. 5.18, the modification due to non-radiative (absorptive) coupling is shown by the dashed curve. We may observe that, at short distances, it is non-radiative coupling that dominates. Moreover, for a given distance, the relative weights of radiative and non-radiative coupling depend on the orientation of the emitting dipole moment.

This simple example illustrates the main features of dipole emission in the vicinity of a nanostructure: the simultaneous existence of radiative and non-radiative coupling, whose relative weights depend on the distance and orientation of the transition dipole moment of the atom, with non-radiative coupling dominating at short distances. It is these features that make the observation of fluorescence from single atoms and molecules a complex and attractive field of study in near-field optical microscopy, both experimentally and from a modelling point of view (see, for example, [14] and Chaps. 12 and 13 of [15]).

## References

1. Born, M., and Wolf, E.: *Principles of Optics*, 6th edn., Cambridge University Press, Cambridge (1980)
2. Bowman, J.J., Senior, T.B.A., and Uslenghi, P.L.E.: *Electromagnetic and Acoustic Scattering by Simple Shapes*, North-Holland, Amsterdam (1969)
3. Carminati, R.: *Modélisation de la microscopie optique de champ proche*, Chap. 2 of [5]
4. Chance, R.R., Prock, A., and Silbey, R.: Molecular fluorescence and energy transfer near interfaces, *Adv. Chem. Phys.* **37**, 1–65 (1978)
5. Chartier, G.: *Manuel d'Optique*, Hermès, Paris (1997)
6. Courjon, D., Vigoureux, J.-M., Spajer, M., Sarayeddine, K., and Leblanc, S.: *Appl. Opt.* **29**, 3734 (1990)
7. de Fornel, F., Goudonnet, J.-P., Salomon, L., and Lesniewska, E.: *ECO2 Opt. Storage Scanning Technol.* **1139**, 77 (1989)

8. Draine, B.T.: The discrete dipole approximation and its application to interstellar graphite grains, *Astrophys. J.* **333**, 848–872 (1988)
9. Greffet, J.-J., and Carminati, R.: Image formation in near-field optics, *Prog. Surf. Sci.* **56**, 133–237 (1997)
10. Girard, C., and Dereux, A.: Near-field optics theories, *Rep. Prog. Phys.* **59**, 657–699 (1996)
11. <http://bernstein.harvard.edu/XieHome.html>
12. <http://www.u-bourgogne.fr/LPUB/opsup/Aaoptisub/Aaoptisubse4.html>  
x7-160004.4
13. Jackson, J.D.: *Classical Electrodynamics*, John Wiley, New York (1975)
14. Klimov, V.V., Ducloy, M., and Letokhov, V.S.: Spontaneous emission of an atom in the presence of nanobodies, *Quantum Electr.* **31**, 569–586 (2001)
15. Courjon, D., and Bainier, C. (Eds.): *Le Champ Proche Optique: Théorie et Applications*, Collection Scientifique et Technique des Télécommunications, Springer-Verlag, Paris (2001)
16. Perez, J.P., Carles, R., and Fleckinger, R.: *Electromagnétisme: Fondements et Applications*, Masson, Paris (1996)
17. Rahmani, A., and de Dornel, F.: Emission Photonique en Espace Confiné, *Le Champ Proche Optique: Théorie et Applications*, Collection Scientifique et Technique des Télécommunications, Eyrolles, Paris (2000)
18. Rosencher, E., and Vinter, B.: *Optoélectronique*, Masson, Paris (1998)
19. Synge, E.H.: *Philos. Mag.* **6**, 356 (1928)
20. Van Bladel, J.: *Singular Electromagnetic Fields and Sources*, Oxford University Press/IEEE (1991/1995)
21. Van Labeke, D., and Barchiesi, D.: Probes for scanning tunneling microscopy: A theoretical comparison, *J. Opt. Soc. Am. A* **10**, 2193–2201 (1993)
22. Vigoureux, J.M., and Courjon, D.: Detection of nonradiative fields in the light of the Heisenberg uncertainty principle and the Rayleigh criterion, *Appl. Opt.* **31**, 3170–3177 (1992)
23. Vigoureux, J.M., Depasse, F., and Girard, C.: Superresolution in near-field optical microscopy defined from properties of confined electromagnetic waves, *Appl. Opt.* **31**, 3036–3045 (1992)

---

## Emerging Nanolithographic Methods

Y. Chen and A. Pépin

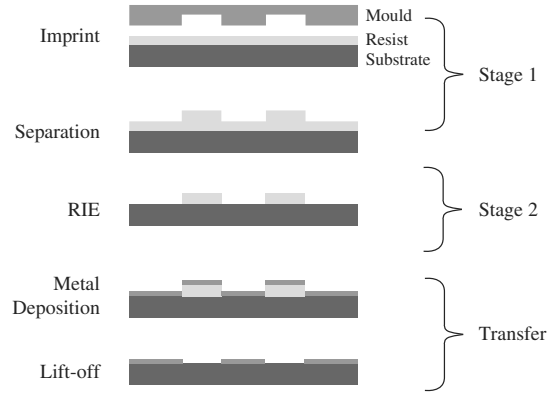
In parallel with the techniques discussed in earlier chapters, several novel, lower cost methods have been developed to provide easier and faster access to the various fields of nanoscience and nanotechnology. This chapter explains the basic principles, performance and fields of application of these emerging methods.

### 6.1 Introduction

Today, the microelectronics industry can produce transistors with features of critical size close to 100 nm on 8-inch wafers. By developing new lithographic methods, such as extreme UV lithography, it will be possible to achieve a critical dimension around 20 nm [35]. To reach still smaller sizes, a further technological breakthrough will be required.

Nanolithography also plays a key part in fundamental research and many areas of industry [7]. The problem is that the methods developed by the electronics industry are often largely inaccessible to research and development teams. This explains the growing interest in cheaper and more flexible methods, as provided by certain emerging nanolithographic methods. These are not based on the conventional use of UV or X-ray photons, or charged particles such as electrons or ions, because the aim is to avoid all the problems of diffraction or scattering. Instead, they use moulding or casting to form high resolution surface patterns on a substrate. The problem of diffraction or scattering usually encountered in conventional lithography will not be an issue and the resolution of all emerging nanolithographic methods is potentially extremely high.

In this chapter, we shall examine several examples of these methods: nanoimprint lithography, nanoembossing, soft lithography, and near-field lithography. We shall also consider the fields of applications of these methods.



**Fig. 6.1.** Nanoimprint lithography: the imprint stage consists in moulding a polymer layer deposited on a substrate; the etch stage removes the residual polymer layer right down to the substrate to create a suitable profile for pattern transfer. The transfer method generally known as lift-off, commonly used in research laboratories, is also illustrated

## 6.2 Nanoimprint Lithography

Nanoimprint lithography is a two-step process (see Fig. 6.1):

- patterns are stamped using a mould on a polymer (resist) layer deposited on the substrate;
- the resist relief pattern stamped onto the polymer is treated by reactive ion etching (RIE) until its recessed areas are all removed.

The final profile of the polymer resist is perfectly comparable with results obtained by other more conventional lithographic methods. This makes nanoimprint lithography compatible with transfer methods commonly used in research establishments. Figure 6.1 also shows the transfer method known as lift-off, as an example. A metal film is deposited and the resist is then dissolved, leaving a very high resolution metal pattern on the substrate.

Nanoimprint lithography was first suggested by S.Y. Chou in 1995 [12,13]. It soon became a standard technique by virtue of its very high resolution and its aptitude for high throughput production. Moreover, this is a simple and cheap method, easy to implement, accessible and applicable in a wide range of situations.

The mould is typically produced by electron beam lithography on a silicon oxide substrate, followed by reactive ion etching (RIE) or electrodeposition [12,19]. To facilitate separation after moulding, the mould surface can be modified by an anti-adhesive treatment.

Thermoplastic polymers such as polymethylmethacrylate (PMMA) or polycarbonate (PC) are often used as imprint resists. The polymer is first dissolved in a solvent, then deposited on the substrate by centrifugation using a spin coater. A soft bake, well above the glass transition temperature  $T_g$  of



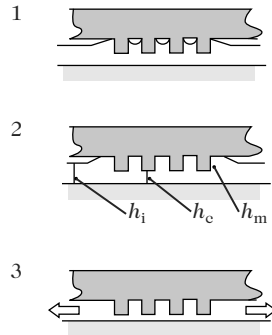
the polymer, serves to evaporate the solvent. Imprinting is also carried out at a temperature above  $T_g$ , at a pressure of a few tens of bar for PMMA, during several minutes. Finally, the system is cooled whilst maintaining the pressure. Once the temperature has fallen below  $T_g$ , the pressure is released and the mould separated from the sample (stage 1 in Fig. 6.1). The relief imprinted in the polymer layer is then transferred to the substrate by RIE, whereupon the residual layer can be removed (stage 2 in Fig. 6.1).

At first sight, the initial step in the process of nanoimprint lithography would appear very similar to the technique of hot embossing, i.e., a polymer layer is thermally deformed using a rigid mould. However, the very small feature sizes to be imprinted and the difficulties related to the material transportation of a viscous polymer in a very restricted space imply exceptional rheological conditions. A fundamental understanding of these phenomena would be extremely useful, but little progress has yet been made with this problem. In practice, an empirical approach provides quick answers to the question of which imprinting parameters to adopt.

A thermoplastic polymer becomes soft and melts above  $T_g$ , with a very low Young's modulus. Cooling below  $T_g$  then fixes the applied deformation, maintaining the shape imposed at the higher temperature. When a polymer has melted, it can be considered as a perfectly viscous fluid [24]. The deformation rate is thus proportional to the shear stress or the applied pressure, and inversely proportional to the viscosity  $\eta$  of the fluid. Now  $\eta$  and  $T_g$  increase with the molecular weight of the polymer. It is therefore better to use polymers with low molecular weight. Note that the viscosity of a polymer is highly sensitive to the temperature. At sufficiently high temperatures, the viscosity begins to decrease exponentially with the temperature. One therefore works at the highest possible temperature at which the polymer does not actually decompose.

The moulding stage of nanoimprint lithography moves polymer around from one place to another. It is clear that the imprinting speed can be increased if the displacement of polymer is limited. Moreover, it is also useful to minimise the residual thickness  $h_c$  remaining at the end of the imprint stage, so that the RIE transfer does not significantly alter the imprinted relief (see Fig. 6.2). Neglecting the change in volume of the polymer, the conservation of matter can be expressed by a relationship between the thickness  $h_m$  of mould features (the depth of features etched into the mould), the initial polymer thickness  $h_i$ , and the residual thickness  $h_c$  of those features cleared out of the polymer. In the case of periodic structures, this requires the relationship  $h_i = h_c + fh_m$ , where the factor  $f$  is the ratio between the etched area of the mould and the total area of the mould, i.e., the etched proportion of the mould surface [9].

This equation represents the condition for an optimal imprint. When the polymer thickness  $h$  under the protruding features of the mould is greater than  $h_c$ , the cavities in the mould will not be completely filled up. When  $h \leq h_c$ , the cavities are completely filled and deformation becomes very slow,



**Fig. 6.2.** Three imprint regimes: (1) incomplete moulding, (2) optimal moulding, (3) over-moulding (mould forced down too far). In order to work in the second regime, the right relationship must be found between the depth  $h_m$  of the mould features, the initial thickness  $h_i$  of the polymer, and the residual thickness  $h_c$  of those features that have been cleared out. In the case of periodic structures, for example, the factor  $f$  represents the ratio between the area of the etched part and the total area of the mould

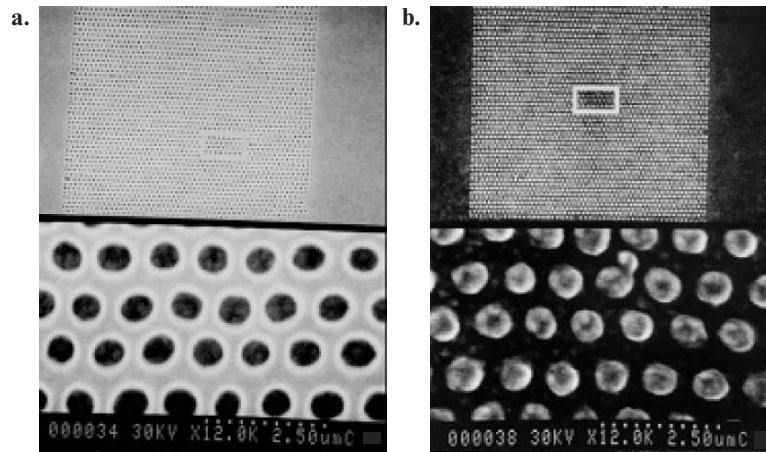
because the matter under the mould must now be forced out at the sides of the mould. With  $h_m = 150$  nm and  $h_c = 20$  nm, we obtain  $h_i = 95$  nm for a line grid, 140 nm for a dot array, and 50 nm for a hole array. It is clear that the initial thickness of polymer that should be deposited depends sensitively on the types of pattern to be imprinted. For arrays of different dimensions or different geometries, one must choose the greatest value of  $f$  for a perfect imprint. However, it is desirable to distribute the features uniformly over the whole surface of the mould, keeping the spacings large enough to serve as reservoir zones.

The minimum feature size that can be achieved by nanoimprint lithography should be much smaller than the molecular size of the polymer, and on this scale, the rheological behaviour can be very different from what is observed on a macroscopic scale. Recall that a polymer is a macromolecule made up of a set of identical molecular chains in different conformations. If the polymer is considered as a random coil, its radius of gyration  $R_g$  can be significantly greater than the dimensions of the features to be imprinted. A theoretical study is therefore in order. Experimentally, the nanoimprint processes have been studied by considering a number of factors including mould fabrication, choice of polymer, temperature, pressure and duration of moulding, and etch parameters as well as specifications of various applications [25].

In nanoimprint lithography, the thickness of imprinted features is very limited when high resolution is required. In order to obtain a high aspect ratio, i.e., the ratio between the thickness and the width of a feature, a new process has been developed involving three layers (see Fig. 6.3). An upper layer of PMMA is imprinted and the pattern transferred by RIE onto a lower supporting layer (e.g., PMGI) which is heat stable at the imprint temperature,



**Fig. 6.3.** Three-layer nanoimprint process invented to obtain high aspect ratios. Fabrication consists in imprinting the upper layer (PMMA), then transferring imprinted features by RIE to the relatively thick underlying resist layer (in this case, PMGI), which is heat stable at the imprint temperature, via a thin intermediate metal film (here, Ge)



**Fig. 6.4.** Dot array with period 100 nm, obtained by trilayer nanoimprint before (a) and after (b) the metal (Ni) lift-off stage

via a thin intermediate metal film (e.g., Ge). With this trilayer process, the following has been achieved:

- routine fabrication of dot arrays with period 100 nm (see Fig. 6.4),
- critical size control close to 10 nm,
- lines etched in silicon (Si) with widths below 10 nm (see Fig. 6.5),
- imprinting at room temperature and/or at low pressure,
- pattern transfer for a variety of applications (see Sect. 6.3).

Furthermore, the homogeneity of the imprint has been demonstrated on a 4-inch wafer, and the pattern positioning accuracy was better than 30 nm for an area of  $30 \times 30 \text{ mm}^2$  [25].

This method of nanoimprint lithography has a potentially very high throughput, since it is a parallel process. However, owing to the heating and cooling stages, the total imprint time easily exceeds a few minutes, well outside the standards required by the world of industry. To speed up the replication

process, another technique has been proposed, i.e., UV-assisted nanoimprint lithography, which consists in moulding, at room temperature and low pressure, a viscous solution of monomer and catalyst mixture, and then solidifying it under UV irradiation through a transparent template. We shall discuss this method in the next section.

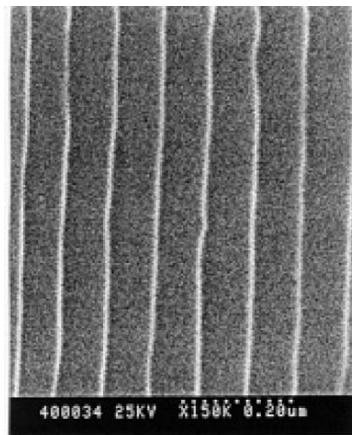
Finally, the fine alignment of nanoimprint lithography is one of the most critical issues for multilevel device fabrication. By using an optical aligner, an alignment accuracy of the order of  $1\ \mu\text{m}$  has been obtained [47], but it is difficult to reach a much higher alignment accuracy because of the difficulties of high temperature and high pressure resist processing. In practice, fabrication of multilevel components can be achieved by combining nanoimprint lithography for the most critical level and optical lithography for other levels.

### 6.3 Applications of Nanoimprint Lithography

The field of application of nanoimprint lithography covers almost all areas requiring high-resolution lithography. We shall discuss various examples in this section.

#### 6.3.1 Microelectronics

S.Y. Chou and coworkers were soon able to fabricate MOSFET transistors by nanoimprint lithography and RIE on an SOI layer (silicon-on-insulator) [15]. More recently, high-frequency low-noise transistors have been made by imprinting T-shaped gates, followed by electrodeposition or lift-off, on GaAs/AlGaAs heterostructures [7]. Combining nanoimprinting with optical

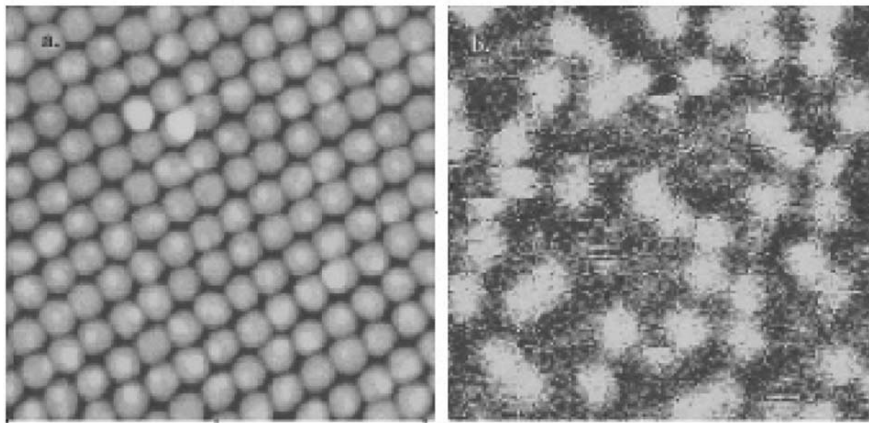


**Fig. 6.5.** Array of silicon lines with width less than 10 nm, obtained by nanoimprint lithography and reactive ion etching of an SOI substrate

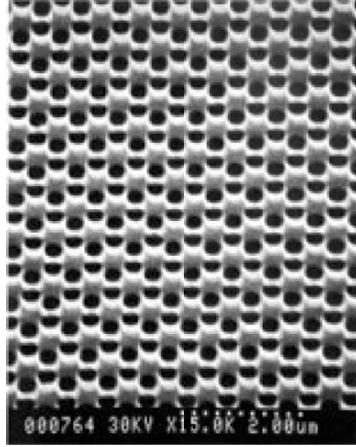
lithography, Förchel and coworkers in Germany have made quantum dot transistors [26]. By virtue of the simple way it creates structures, nanoimprinting has also been used recently to fabricate organic transistors [1] and high-density molecular memories [9]. Figure 6.5 shows an Si line array in which the lines have widths less than 10 nm, made by nanoimprinting and etching an SOI substrate. With these very fine lines, it should be possible to make extremely high performance Si transistors.

### 6.3.2 Nanomagnetism

Historically, nanoimprint lithography arose from research into new ways of fabricating magnetic disks from nickel columns in a silicon matrix [14]. Nanoimprinting is used to define patterns on a silicon substrate. After imprinting, a thin metal film is deposited on the upper surface of the resist (oblique angle deposition), followed by RIE of the silicon and an electrolytic deposition. The process ends with chemomechanical polishing, leaving a plane surface covered with high density magnetic structures [22]. Nanoimprinting has also been used to fabricate spin-valve structures [21], emphasising the crucial importance of the lift-off parameters and the RIE during the pattern transfer stage. By using nanoimprinting followed by lift-off, several types of nanostructure have been studied in detail: multilayer Pt/Co dots magnetised perpendicularly to the plane, magnetic dots and rings of polycrystalline cobalt magnetised parallel to the plane, with different lateral sizes, thicknesses, and spacings [9]. Figure 6.6 shows an atomic force microscopy (AFM) image (left) and a magnetic force microscopy (MFM) image (right) of a magnetic dot (Co) array with period



**Fig. 6.6.** Array of magnetic dots (Co) with period 60 nm (storage density 180 Gbit/in<sup>2</sup>), made by nanoimprint lithography and lift-off. *Left:* Atomic force microscopy (AFM) image. *Right:* Magnetic force microscopy (MFM) image. Courtesy of B. Diény, CEA, Grenoble

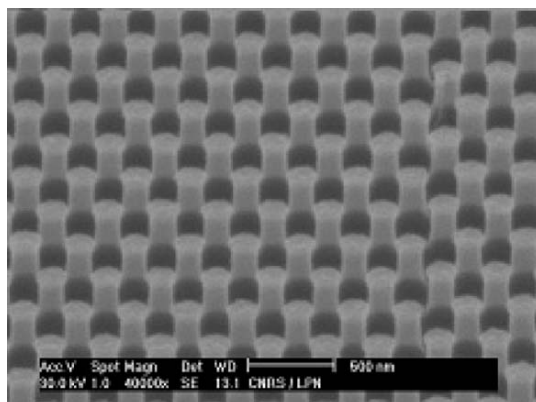


**Fig. 6.7.** Graphite-type array on a polymer layer, for use as an etch mask for the underlying substrate in order to obtain functional photonic crystals

60 nm made by nanoimprint lithography and lift-off [31]. More recently, the IBM team at Almaden has produced magnetic nanostructures with period 100 nm over a large area by UV nanoimprinting with a flexible template [27].

### 6.3.3 Nano-Optics

Much has been achieved in this field for passive and active optical devices. Examples are the fabrication of metal–semiconductor–metal photodetectors, grating polarisers with period 190 nm, waveguide polarisers, planar resonators, infrared filters, photonic crystals, and antireflective layers [38]. Nanoimprinting can also be used to introduce spontaneous anisotropy of molecules or chromophores during imprinting [42]. Patterns have been replicated on a nano-compact disk with a density of 400 Gbit/in<sup>2</sup> by nanoimprinting without [23] and with [39] UV irradiation. Active optical devices have been fabricated by directly moulding diffraction gratings into a polymer layer and then depositing dye molecules on them, or directly moulding a polymer layer already doped with emissive molecules. More recently, a feasibility study has been carried out concerning the fabrication of high density organic light-emitting arrays [11]. Finally, nanoimprint lithography seems to be a viable technique for duplicating photonic crystals, optical nanostructures widely considered to be the building blocks for a new optics. Figure 6.7 shows a graphite array fabricated in a polymer layer, which can be used as a mask for etching the underlying substrate to obtain GaAs photonic crystals, for example.



**Fig. 6.8.** Nanopillar array integrated into a microfluidic channel to improve the separation of DNA molecules by capillary electrophoresis on a chip

#### 6.3.4 Chemistry and Biology

Nanoimprint lithography has also been used in several applications to chemistry and biology [32]:

- selective modification of the surface properties of a substrate by silanisation [34],
- the fabrication of biosensors in the form of a high density interdigital electrode array [30],
- the fabrication of nanostructures for separating DNA on a microfluidic chip [32].

The duplication of nanostructures for sorting DNA molecules is of special interest to biologists. Figure 6.8 shows an array of nanopillars integrated into a microfluidic channel, made by nanoimprint and RIE on a silicon oxide wafer. Combining nanoimprint lithography with a grazing-incidence deposition of silicon, Chou and coworkers have obtained nanochannels of various dimensions [4]. By controlling the angle and thickness of the deposit, channels as small as 10 nm have been achieved. Another example is the fabrication of a surface with topography designed to carry out cell engineering [45]. It has been shown that cells start to grow along the edges of etched structures and do not adhere on a surface very densely covered with pillars, a situation which may prove useful for the repair of tendon tissues.

Generally speaking, nanoimprint lithography is very useful when fabrication involves the reproduction of nanostructures over a large area but without requiring too high a level of complexity, e.g., accurate alignment.

## 6.4 UV Nanoimprint Lithography (UV-NIL)

UV nanoimprint lithography consists in moulding at room temperature a viscous pre-polymer made of a monomer (or pre-polymer) and catalyser mixture, and then photopolymerising it, i.e., solidifying it under UV irradiation through a transparent template (see Fig. 6.9a). The first experiment to achieve genuine nanoscale resolution was published by Haisma and coworkers in 1996 [20]. The advantage of this technique lies in the fact that it works at room temperature and low pressure using a quartz template. It also offers a high duplication rate (high throughput) and the possibility of high precision alignment.

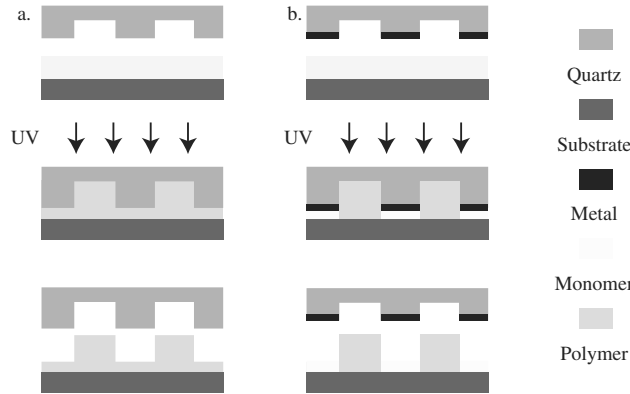
C.G. Wilson and team at the University of Texas (Austin) used the same basic idea to develop a new process called step-and-flash [17]. In this process, a small template is used to pattern a large-area wafer. The template is first brought down towards the silicon wafer at a well-defined location. The mixture of monomer solution and catalyst is injected into the space between the template and the wafer. The solution spreads out due to capillary forces (surface tension). A slight pressure is applied and the part of the solution situated under the template is polymerised locally by UV irradiation. This process is repeated over the whole wafer to obtain a homogeneous imprint on the surface. After imprinting, the residual polymer layer is removed by RIE and the ensuing process is similar to thermal nanoimprint lithography. Resolutions better than 100 nm have been obtained in this way over a large area.

Compared with standard nanoimprint lithography, UV imprinting can be carried out at room temperature and lower pressures (of the order of 1 bar). It has also been demonstrated that this technique can be used to imprint the same patterns at different positions on a silicon wafer, which is extremely useful for the purposes of mass production. Another advantage with this method is that the template is self-cleaning by virtue of the photopolymerisation process at each imprint step [18].

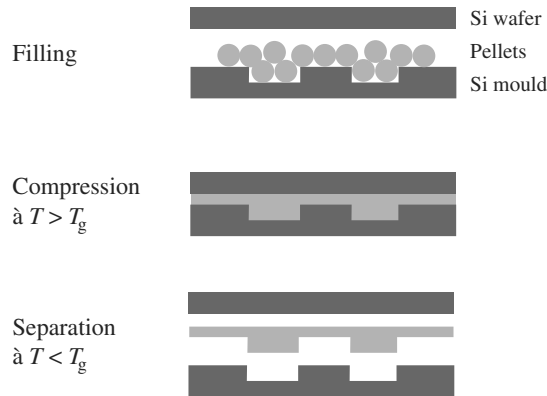
H. Kurz and coworkers at the University of Aix-la-Chapelle in Germany [3] has investigated another process using a sol-gel resist of very low viscosity. In order to improve separation of the template and sample, 1–2% of fluorinated molecules were added, thereby allowing homogeneous patterning over a large area. To achieve this, semi-rigid transparent templates can be used, consisting in a top imaging layer, an elastomeric buffer layer allowing large area conformability, and a quartz carrier.

The UV imprint process can also be improved by considering new template configurations. One such configuration consists of a top imaging layer on which the non-etched surface is coated with a thin absorbing film, i.e., a material absorbing UV radiation, such as a metal (see Fig. 6.9b). A mask of this kind can be used to form a near-field image with good optical contrast around etched features, and hence to replicate features as small as 50 nm by a photopolymerisation process [6].





**Fig. 6.9.** UV nanoimprint lithography: (a) with an etched quartz template, leading to photopolymerisation of the monomer everywhere beneath the template; (b) with a quartz template partially coated with a metal film, resulting in a partial polymerisation making it easier to remove the non-polymerised part



**Fig. 6.10.** Nanoembossing process using pellets of a thermoplastic polymer and a high resolution etched silicon mould

### 6.5 Nanoembossing

Nanoembossing is a technique for forming nanostructures on the surface of a bulk material. It can be used to make functional nanostructures or all-plastic devices. Three approaches have been studied:

- direct imprint on the surface of a plastic wafer [2] at a temperature below  $T_g$  and a relatively high pressure so as not to melt the wafer;
- imprint of nanostructures on a polymer film deposited on a rigid substrate, without a subsequent etching stage;
- compression of thermoplastic polymer pellets at a temperature above  $T_g$  to form a nanopatterned wafer.

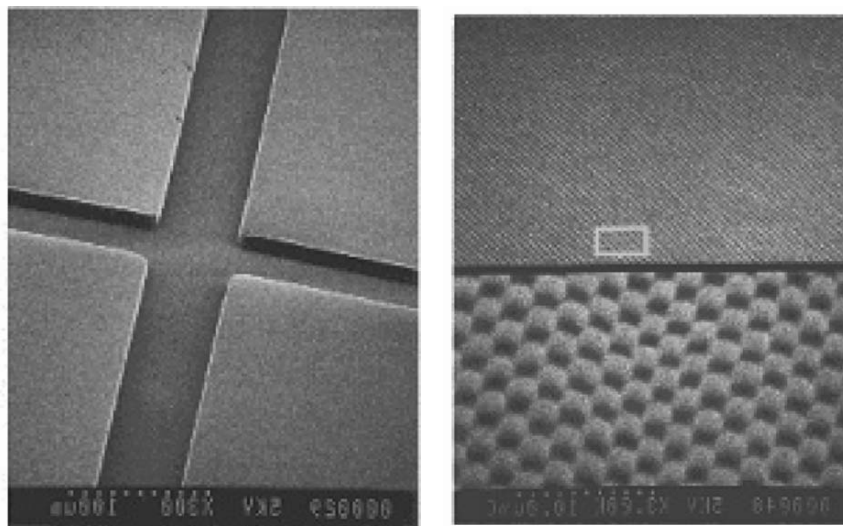
Figure 6.10 illustrates the basic idea underlying the latter technique. A mould is first fabricated in Si by electron beam lithography and RIE. Several plastic pellets are placed between the mould and a planar Si wafer. This sandwich is then positioned between the two hot plates of a press. When the temperature goes above  $T_g$  for the polymer, a high enough pressure is applied to compress the melted pellets. After cooling below  $T_g$ , the imprinted plastic sheet is detached from the mould. This method is very low cost and very high resolution, since it uses pellets melted in the same way as in injection moulding and an Si mould etched on the nanoscale as in nanoimprint lithography.

Whatever the size and geometry of pattern features, nanoembossing can reproduce them to high accuracy. In particular, it can be used to make deep microchannels and shallow nanostructures in a single step. To obtain structures combined on a single Si mould, a new two-step process (optical lithography and electron lithography) has been developed. Optical lithography has been used to expose a thick resist (commercially available SU-8), with thickness  $10\ \mu\text{m}$ , using a mask defining microfluidic channels. After development, the patterns in the SU-8 are used as a mask to etch the Si by RIE with a fluorinated plasma ( $\text{SF}_6$ ). Shallow nanostructures have been obtained by electron beam lithography, followed by lift-off (Ni) and RIE [37].

Nanoembossing has been investigated using different types of polymer. For PMMA, the embossing temperature has been fixed at  $180^\circ\text{C}$ , as for nanoimprint lithography. To obtain a thin sheet ( $\sim 100\ \mu\text{m}$ ) with diameter 2 to 3 inches, 4 or 5 PMMA pellets is sufficient. To obtain thicker wafers, a metal frame has been used to confine the pellets laterally. Figures 6.11a and b show a microfluidic channel  $10\ \mu\text{m}$  deep and a nanostructure array with period  $300\ \text{nm}$  integrated into the channel, designed to separate DNA or protein molecules by capillary electrophoresis. A critical stage occurs when this kind of microchannel is closed. It is important to seal the system whilst being careful not to block the nanopatterned channel. After studying several bonding methods, thermal bonding turned out to be the best solution. This involves pressing together two plastic wafers with a low pressure at a temperature close to  $T_g$  for the polymer.

Plastic materials are used because of their low cost and biocompatibility. Nanoembossing has a wide field of application, ranging from the fabrication of diffracting optical elements to microfluidic devices. All-plastic microfluidic systems can be fabricated at relatively low costs and with nanoscale resolution using this technique. It is also suitable for making hybrid devices with various functionalities. Regarding the medium and long term prospects, it should be possible to make active elements for mechanics, optics and electronics based on nanoembossing of polymers with novel characteristics.

Nanoembossing can be used to pattern other types of material. For example, Chou and coworkers have recently been able to imprint nanostructures directly onto a silicon wafer [16]. To do so, they melt a thin film of Si on the surface of a wafer by shining an excimer laser through a quartz template. A single 20-ns laser pulse suffices to melt the Si to a depth of several hundred



**Fig. 6.11.** Microchannel (*left*) and nanostructures (*right*) made by nanoembossing with PMMA pellets and an etched silicon mould

nanometers. Under pressure, the melted Si, in the liquid state, is easy to shape. After the pulse, the melted Si solidifies very quickly, leaving a perfect copy of the pattern on the template. The whole process, including heating, stamping and cooling, only takes 250 ns. To illustrate the duplication capabilities of this method, an etched grating of period 300 nm and depth 110 nm has been stamped, accurately reproducing features down to 10 nm. The same embossing process has been used to stamp large slabs and thin films of polycrystalline Si deposited on an Si substrate coated with a 200-nm silica layer, as well as metal films [16]. Naturally, this innovation has inspired a great deal of interest in the nanoembossing technique.

## 6.6 Soft Lithography

Soft lithography covers a range of techniques initially proposed by G.M. Whitesides at the beginning of the 1990s. These techniques are based on the use of an elastomer, polydimethylsiloxane (PDMS), as a mould or ink stamp [32, 43, 46]. PDMS is a silicone oil polymer formed by repetition of the monomer  $-\text{OSi}(\text{CH}_3)_2\text{O}-$ , used commercially to produce flexible contact lenses, among other things. This type of polymer material is commonly referred to as soft matter.

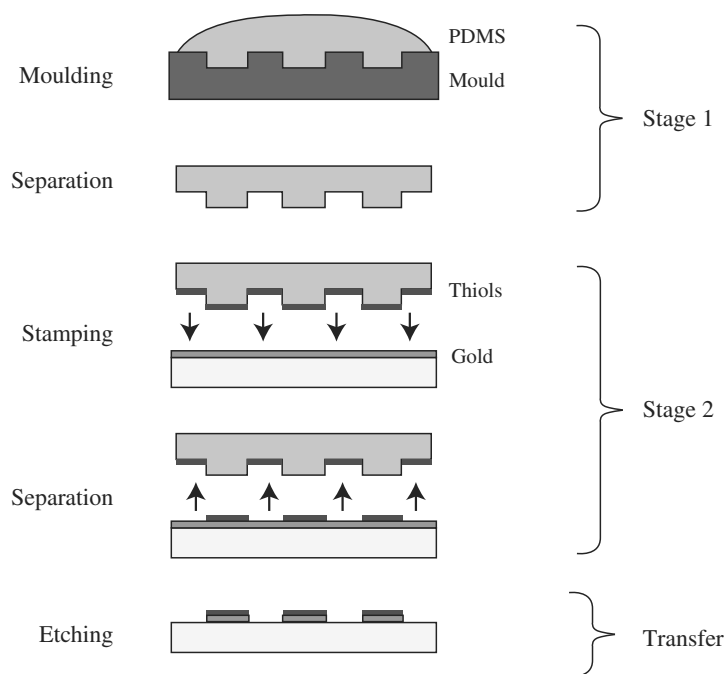
To produce a replica in PDMS by soft lithography, the first step is to make an initial mould, or master, out of a resist or some rigid material, such as silicon or a metal. When nanoscale resolution is required, this master is fabricated

by some form of high resolution lithography, such as electron beam lithography and reactive ion etch techniques. A liquid mixture of two commercially available components is then poured onto the master: one is a chemical precursor of PDMS and the other a crosslinking catalyst which favours crosslinking between molecules. This reaction is accelerated by heating, typically for one hour at 80°C, and the solidified elastomer is then detached from the mould (see stage 1 of Fig. 6.12). To facilitate separation after moulding, the mould surface can be given a preliminary anti-adhesive treatment. The degree of crosslinking and elastic properties of the PDMS, such as its Young's modulus, can be adjusted by modifying the initial proportions of the two components and/or by adding specific chemical agents to the mixture. PDMS is optically transparent, chemically inert and mechanically deformable. Applications of PDMS-based soft lithography include [44]:

- microcontact printing,
- moulding resist patterns using capillary forces,
- transplanted matter or deposition of (bio)molecules via cavities in the mould,
- fabrication of microfluidic devices.

In particular, microcontact printing (or stamping) can be used to print nanoscale patterns onto a solid surface (see stage 2 of Fig. 6.12). The patterned PDMS stamp is dipped into a solution of organic molecules, e.g., thiols in ethanol. It is then applied to a gold film deposited on a substrate. The sulfur-bearing extremities of the thiols allow them to adsorb onto the gold, whereupon they form a self-assembled monolayer reproducing the pattern on the PDMS stamp. The regions that have come into contact with the stamp are protected by the thiol monolayer and resist wet chemical etching, so that the pattern can be transferred to the gold film, then to the substrate, if necessary. By first depositing a resist layer, e.g., PMMA, between the substrate and the gold film, the microcontact printing method can be made compatible with standard lithographic processes, since a simple RIE of the PMMA using the chemically etched gold patterns as mask can create resist profiles that are perfectly suited to the lift-off process.

The main advantage of microcontact printing is the possibility of patterning large areas in a single step by means of a large stamp, or indeed a rotating stamp. Moreover, as these flexible stamps can deform to fit the substrate topography, this technique can also be used to pattern curved surfaces. Other organic molecules than the thiols, e.g., trichlorosilanes, proteins, etc., can be stamped onto different surfaces, e.g., silicon oxide, glass, polymers. Although this moulding technique can reproduce features of the PDMS replica with an accuracy of a few nanometers, the resolution obtained by microcontact printing is limited by diffusion of the printed molecules, both before and after contact (the ink tends to run), but also by the deformation and distortion suffered by the elastomeric stamp. However, resolutions of 50 nm have been achieved in optimal inking conditions [28].

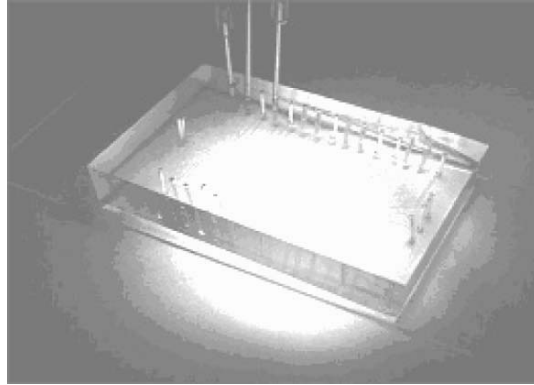


**Fig. 6.12.** Soft lithography (stage 1) and microcontact printing (stage 2). Soft lithography consists in polymerising a mixture of a monomer and a catalyst in a master at a temperature of  $\sim 80^\circ\text{C}$  for one hour. After separation, the solidified elastomer can be used as an ink stamp to transfer self-assembled molecules onto a substrate. The pattern is then transferred onto the substrate by chemical etching

This low cost micro/nano-printing technique has a wide field of applications, notably in biochemistry and biology. For instance, it has recently been used to deposit an array of single proteins with 80-nm resolution, using a PDMS stamp with high Young's modulus, hence rather rigid [28]. However, as for nanoimprint lithography, it remains a challenge to use this method to make complex devices requiring high-precision alignment between successive lithographic levels, as happens in the production of microelectronic circuits.

The moulding of resist structures via capillary forces can also be used to define structures with resolutions of the order of 10 nm. The PDMS stamp is then placed on a rigid surface and a liquid polymer enters the cavities of the mould by capillary forces (surface tension). The polymer subsequently solidifies according to the mould patterns. The channels thereby formed, known as microfluidic networks, can also be used for local deposition of molecules such as proteins, or certain catalysts and chemical reagents.

Due to its low cost, simplicity, rapidity, and the properties of PDMS, the fabrication of microfluidic devices by soft lithography is now widespread. The most common approach is to mould the microchannels in a layer of PDMS using a micron-scale resist mould obtained by conventional optical



**Fig. 6.13.** Multilevel microfluidic device made from PDMS using soft lithography. The device integrates microchannels, micropumps, microvalves and connecting elements for applications in cell and molecular biology

lithography. The PDMS layer is subsequently pierced with a needle at the appropriate points so that liquids can be injected into the channels. The device is then sealed against a glass plate after an oxygen plasma surface treatment, to form a microfluidic circuit. This method can be used as a quick way of carrying out biochemical analyses like capillary electrophoresis.

Nanochannels can be obtained using moulds with nanoscale resolution. Multilayer devices containing several levels of channels – either independent or connected together to form a 3D fluid network – can also be fabricated quite simply by preparing several PDMS replicas separately and then superposing them on a glass plate. S.R. Quake and coworkers at Caltech have developed a novel multilayer process exploiting the deformation properties of PDMS and a crisscross architecture of channels on two levels separated by a thin PDMS membrane that can easily be deformed with pressure to close one of the channels. This provides a simple way to make high performance pneumatic microvalves and micropumps (see Fig.6.13) [40]. These microfluidic devices have already proved their efficacy for sorting and handling cells, sorting DNA molecules, parallel crystallisation of proteins and other functions. However, the fabrication of nanoscale elements using this process is still under study.

Finally, by combining soft lithography with other nanofabrication methods, either conventional or non-standard, other functional units have been devised for use in both fundamental and applied research.

## 6.7 Near-Field Lithography

Among the non-standard replication methods, it is worth mentioning near-field optical lithography, which consists in exposing a thin film of resist through a mask in perfect contact with the substrate. Traditionally, contact

optical lithography uses quartz masks and perfect contact is frustrated by the rigidity of both mask and substrate (Si). By approaching perfect contact, masks made from PDMS or some other deformable configuration are able to achieve resolutions of around 100 nm by optical lithography. Indeed, by using a totally transparent PDMS mask (without absorbers), the phase effect leads to the formation of near-field images with an optical contrast that is sufficient to expose a thin resist layer [33]. This method has been improved by the IBM group in Zurich by introducing a metal film into the part of the mask that is not in contact, thereby enhancing the image contrast by virtue of the coupling between the PDMS structures and the substrate [36]. The advantage with these techniques lies in their simplicity and the flexibility of mask fabrication. There have been applications in micro-optics [33] and micromagnetism [9].

Near-field techniques also include those methods wherein a tip is used to scan a surface, such as scanning tunneling microscopy (STM), atomic force microscopy (AFM), and scanning near-field optical microscopy (SNOM). These techniques are described in detail in Chaps. 3–5, so we shall only mention their applications to nanofabrication very briefly here. Using the tunnel current between an STM tip and the substrate, a very small amount of matter can be deposited on the substrate, or the substrate can be etched locally. Many experiments have been carried out under a wide range of conditions, e.g., room temperature, low temperature, ultrahigh vacuum, etc., achieving resolutions well below 100 nm.

Using a scanning tunneling microscope, it is also possible to displace atoms one by one to fabricate patterns on a surface and, in certain cases, to design simple electronic devices involving a single molecule. Given that physical and chemical properties are quite different on the atomic scale from those observed on the macroscopic scale, many subjects have been tackled, leading to extremely varied research programmes and applications.

The relatively low write speed of these techniques depends essentially on the scanning frequency. In order to reach reasonable speeds, an array of tips operating in parallel is required. The IBM team in Zurich has developed a technique known as Millipede, which integrates thousands of cantilevers into the same structure [41]. Each cantilever carries an AFM tip and a pair of electrodes for positioning it with very great accuracy on the substrate, so that it can modify the morphology of a polymer layer locally by heating, making it similar in this respect to nanoimprint lithography. This approach was originally suggested for high density storage, but it may also be used as a high resolution lithographic technique when suitably optimised.

Another approach has recently been put forward, known as dip-pen lithography. This method, which associates AFM and microcontact printing, can achieve resolutions of the order of 10 nm. The AFM tip is coated with a solution of organic molecules, usually thiols, which are then deposited locally by self-assembly on a suitably prepared substrate (gold for thiols) [29].

## 6.8 Conclusion

All the alternative nanolithographic methods discussed here aim primarily to fabricate nanostructures and microsystems at low cost. They are particularly useful for designing novel functions that are not in competition with more conventional lithographic methods. These techniques are already widely accepted in several different fields of research. At the present time, however, it is difficult to envisage any large scale industrial implementation. This situation should change as nanolithographic methods become an inescapable feature in the manufacture of certain products of mass consumption.

## References

1. Austin, M.D., and Chou, S.Y.: Appl. Phys. Lett. **81**, 4431 (2002)
2. Becker, H., and Heim, U.: Sensors and Actuators **83**, 130 (2000)
3. Bender, M., Otto, M., Hadam, B., and Spangenberg, B.: Microelectron. Eng. **53**, 233 (2000)
4. Cao, H., and Yu, Z., Wang, J., Tegenfeldt, J.O., Austin, R.H., Chen, E., Wu, W., and Chou, S.Y.: Appl. Phys. Lett. **81**, 174 (2002)
5. Carcenac, F., Vieu, C., Lebib, A., Chen, Y., and Launois, H.: Microelectron. Eng. **53**, 163 (2000)
6. Chen, Y., Carcenac, F., Ecoffet, C., Loughnot, D.J., and Launois, H.: Microelectron. Eng. **46**, 69 (1999)
7. Chen, Y., and Pépin, A.: Electrophoresis **22**, 187 (2001)
8. Chen, Y., Natali, M., Li, S.P., and Lebib, A.: In *Alternative Lithography*, ed. by C.M. Sotomayor-Torres, Kluwer Academic/Plenum Publishers (2003)
9. Chen, Y., Macintyre, D., Boyd, E., Moran, D., Thayne, I., and Thoms, S.: J. Vac. Sci. Technol. B **20**, 2887 (2002)
10. Chen, Y., et al.: Appl. Phys. Lett. **82**, 1610 (2003)
11. Cheng, X., Hong, Y., Kanicki, J., and Guo, L.J.: J. Vac. Sci. Technol. B **20**, 2877 (2002)
12. Chou, S.Y., Krauss, P.R., and Renstrom, P.J.: Appl. Phys. Lett. **67**, 3114 (1995)
13. Chou, S.Y., Krauss, P.R., and Renstrom, P.J.: Science. **85**, 272 (1996)
14. Chou, S.Y.: Proc. IEEE. **85**, 625 (1997)
15. Chou, S.Y., Krauss P.R., and Zhang, W.: J. Vac. Sci. Techn. B **15**, 2897 (1997)
16. Chou, S.Y., Keimel, C., and Gu, J.: Nature **417**, 835 (2002)
17. Coburn, M., et al.: SPIE 24th Intl. Symp. *Microolithography: Emerging Lithographic Technologies III*, Santa Clara, CA, 379 (1999)
18. Bailey, T., Smith, B., Choi, J., Colburn, M., Meissl, M., Sreenivasan, S.V., Ekerdt, J.G., Wilson, C.G.: J. Vac. Sci. Technol. B **19**, 2806 (2001)
19. Gale, M.T.: Microelectron. Eng. **34**, 321 (1997)
20. Haisma, J., Verheijen, M., van dean Heuvel, K., and van den Berg, J.: J. Vac. Sci. Technol. B **14**, 4124 (1996)
21. Kong, L., Pan, Q., Cui, B., Li, M., and Chou, S.Y.: J. Appl. Phys. **85**, 5492 (1999)
22. Krauss, P.R., and Chou, S.Y.: J. Vac. Sci. Technol. B **13**, 2850 (1995)
23. Krauss, P.R., and Chou, S.Y.: Appl. Phys. Lett. **71**, 3174 (1997)



24. Van Krevelen, D.W.: *Properties of Polymers*, Elsevier, Amsterdam (1990)
25. Lebib, A.: Doctoral thesis, Université de Paris 7 (2001)
26. Martini, I., Eisert, D., Kamp, M., et al.: *Appl. Phys. Lett.* **77**, 2237 (2000)
27. McClelland, G., Hart, M.W., Best, M.E., Rettner, C.T., Carter, K.R., and Terris, B.D.: 1st Intl. Conf. on *Nanoimprint and Nanoprint Technology*, San Francisco, December 11–13 (2002)
28. Michel, B., Bernard, A., Bietsch, A., Delamarche, E., Geissler, E., Juncker, D., Kind, H., Renaut, J.P., Rothuizen, H., Schmid, H., Schmidt-Winkel, P., Stutz, R., and Wolf, H.: *IBM J. Res. & Dev.* **45**, 697–719 (2001)
29. Piner, D., Zhu, J., Xu, F., Hong, S., and Mirkin, C.A.: *Science* **283**, 661 (1999)
30. Montelius, L., Heidari, B., Grayczyk, M., et al.: *Microelectron. Eng.* **53**, 521 (2000)
31. Mortitz, J., Landis, S., Nozières, J.P., Lebib, A., Chen, Y., and Dieny, B.: *J. Appl. Phys.* **91**, 7314 (2002)
32. Pépin, A., and Chen, Y.: Chap. 17 in *Alternative Lithography*, ed. by C.M. Sotomayor-Torres (ed), Kluwer Academic/plenum Publishers (2003)
33. Rogers, J.A., and Whitesides, G.M.: *Appl. Opt.* **36**, 5792 (1997)
34. Schift, H., Heyderman, L.J., Padeste, C., and Gobrecht, J.: *Microelectron. Eng.* **61/62**, 423 (2002)
35. *The National Technology Roadmap for Semiconductors*: available at the Website <http://public.itr.net/Files/2002Update/2002Update.pdf> (2003)
36. Schmid, H., Biebuyck, H., Michel, B., Martin, O.J., and Piller, N.B.: *J. Vac. Sci. Technol. B* **16**, 3422 (1998)
37. Studer, V., Pépin, A., and Chen, Y.: *Appl. Phys. Lett.* **241**, 447 (2002)
38. Scheer, H.C., Schulz, H., Hoffmann, T., and Sotomayor Torres, C.M.: *Nanoimprint techniques*, in *The Handbook of Thin Films*, ed. by H.S. Nalwar, Academic Press (2001)
39. Terris, B.D., Manin H.J., Best, E., Logan, J.A., Ruger, D., and Rishton, S.A.: *Appl. Phys. Lett.* **69**, 4262 (1996)
40. Unger, M.A., Chou, H.P., Thorsen, T., Scherer, A., and Quake, S.: *Science* **288**, 113 (2000)
41. Vettiger, P., Despont, M., Drechsler, U., Dürig, U., Häberle, W., Lutwyche, M.I., Rothuizen, H.E., Stutz, R., Widmer, R., and Binnig, G.K.: *IBM J. of Res. & Dev.* **44**, 323 (2000)
42. Wang, J., Sun, X., Zhuang, L., et al.: *Appl. Phys. Lett.* **77**, 166 (2000)
43. Whitesides, G. M., and Love, G.: *Pour la Science*, December 2001, p. 86
44. Whitesides, G.M., and Stroock, A.D.: *Physics Today* **54**, 42 (2001)
45. Wilkinson, C.D.W., Curtis, A.S.G., and Crossan, J.: *J. Vac. Sci. Technol. B* **16**, 3132 (1998)
46. Xia, Y.N., and Whitesides, G.M.: *Angew. Chem. Int. Ed. Engl.* **37**, 550 (1998)
47. Zhang, W., and Chou, S.Y.: *Appl. Phys. Lett.* **79**, 845 (2001)

## **Part II**

---

### **Nanoscale Objects**

## Clusters and Colloids

A. Perez, P. Mélinon, J. Lermé, and P.-F. Brevet

The clusters and colloids discussed in this chapter are structures made up of anything from a few tens to a few thousand or tens of thousands of atoms, with diameters between 1 and 10 nm, which are intermediate states of matter between the molecule and the bulk solid. In this nanoscale world, where the range of interactions is equal to or exceeds the dimensions of the objects, characterised moreover by a large surface to volume ratio, clusters and colloids often manifest novel atomic and electronic structures with unique properties.

It was at the beginning of the 1980s that these systems became the subject of intense study, with the development of supersonic beam techniques capable of producing any kind of cluster, and the advent of soft chemistry in the case of colloids. In parallel, progress in nanotechnology had made it possible to observe and study such small objects, giving a tremendous boost to the field. On a fundamental level, decisive advances were then achieved in understanding the properties of these objects. Examples are provided by quantum size effects in metals, due to the confinement of electrons in a reduced volume, and the effects of geometric structures where the atomic arrangement may differ from the arrangement in the bulk solid and lead to fivefold (icosahedral) symmetry or cage structures like fullerenes. Electronic, optical and magnetic properties may also manifest themselves in spectacular ways depending on the size of the object. For example, surface plasmons confer a different ‘colour’ on clusters than is observed in larger samples, and magnetic moments can vary with the size, to the extent that some non-magnetic materials can exhibit a magnetic moment when the system size decreases far enough. Clusters are also model systems for tackling the question of fragmentation and phase transitions on the nanoscale via evaporation studies, fission, and segregation. We also note the recent interest in the dynamical properties of these tiny systems, right down to the femtosecond scale. Quite generally, the field of clusters and colloids provides a good example of how experiment and theory can work together. In particular, the theoretical aspect has evolved significantly with recent developments in computer methods and tools which allow scientists to model realistic systems of ever increasing size.

In this chapter, we begin by analysing the various parameters and quantities which characterise these kinds of nanoscale objects and condition their specific properties and structures (Sects. 7.1–7.3). Quantum effects typical of these nanoscale systems, such as electronic shell and supershell structures and collective excitation phenomena, are the subject of Sect. 7.4. In Sect. 7.5, we describe the main preparation methods, physical for clusters and chemical for colloids. Finally, in Sect. 7.6, we discuss current approaches used to prepare dense disordered or ordered assemblies of clusters or colloids in order to investigate their specific collective properties for applications in future very high density integrated devices (components, sensors, etc.) for electronics, optics, optoelectronics and magnetism.

For reasons of convenience, we shall often use different terms, sometimes equivalent, in different contexts to refer to the nanometric configurations discussed in this chapter. For example, we may speak of clusters or colloids depending on whether the method of preparation is physical or chemical, whilst the terms nanoparticle, nano-object and nanosystem may be used indifferently for either clusters or colloids when we wish only to specify the nanoscale dimensionality of the object.

The appendix at the end of the chapter provides further information about various aspects of the discussion which can thereby be treated more summarily in the main text.

## 7.1 Equilibrium Shape

The key parameter for calculating the equilibrium shape of a cluster is the cohesive energy of the atoms in a given geometry.

### 7.1.1 Liquid-Drop Model

A droplet (in this case, a cluster) is made up of  $N$  atoms and assumed to be spherical. It will also be assumed that the cluster comprises only one constituent. As we shall see in Sect. 7.2, the properties of a very small object are related to its radius of curvature. This model is transposed here from nuclear physics, where it explains the masses of atomic nuclei, to solid state physics. It does not allow for the object to have facets (plane faces).

The cohesive energy of the  $N$  atoms in this spherical droplet can be written as a volume term reduced by the excess surface free energy. The surface energy is the product of the surface tension  $\gamma$  and the area  $S$ , bearing in mind that there is only one constituent, whence

$$E_{\text{cluster}} = E_{\text{bulk}} - S\gamma, \quad (7.1)$$

where  $E_{\text{bulk}}$  is the total cohesive energy in the bulk solid. If  $r_a$  is the Wigner–Seitz radius of the atom ( $v_0 = 4\pi r_a^3/3$ ), the area  $S$  is

$$S = N^{2/3} 4\pi r_a^2. \quad (7.2)$$

Normalising, it follows that the energy per atom is

$$\frac{E_{\text{cluster}}}{N} = E_{\text{cohesive}} - \frac{4\pi r_a^2 \gamma}{N^{1/3}}, \quad (7.3)$$

where  $E_{\text{cohesive}}$  is the cohesive energy per atom in the bulk, i.e.,  $E_{\text{cohesive}} = E_{\text{bulk}}/N$ . Given that

$$N = \frac{d^3}{(2r_a)^3}, \quad (7.4)$$

where  $d$  is the diameter of the droplet ( $d = 2R$ ), we obtain finally

$$\frac{E_{\text{cluster}}}{N} = E_{\text{cohesive}} - \frac{6v_0\gamma}{d}, \quad (7.5)$$

where the cohesive energy varies as  $1/R$ .

The difficulty with this model is to estimate the ‘area’  $4\pi r_a^2 \gamma$ , which requires knowledge of  $\gamma$ . It has been shown that this term has a value of about 0.82 times the bulk term  $E_{\text{cohesive}}$ . We thus end up with the empirical formula

$$\frac{E_{\text{cluster}}}{N} = E_{\text{cohesive}}(1 - 0.82N^{-1/3}). \quad (7.6)$$

We may define a dimensionless parameter  $E_{\text{norm}}$  which expresses the ratio of the droplet energy in relation to the energy in the bulk:

$$E_{\text{norm}} = \frac{E_{\text{cluster}}}{E_{\text{bulk}}}, \quad (7.7)$$

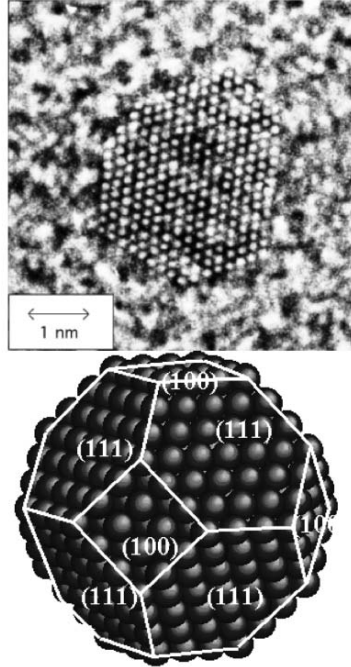
whence

$$E_{\text{norm}} = 1 - 0.82N^{-1/3}. \quad (7.8)$$

We may now predict structural changes on the basis of a thermodynamic approach, using the phase diagram. Indeed, according to the drop model, a pressure is exerted on the cluster, known as the Gibbs pressure, derived from Laplace’s law. This pressure balances the force exerted on the curved solid:

$$P = \frac{2\gamma}{R} \quad (7.9)$$

We shall show later that the melting point decreases as  $1/R$  [see (7.43)]. From a phenomenological standpoint, a cluster may be considered to have the same phase diagram as the bulk solid at high pressure (HP) and high temperature (HT), with a scale factor of  $1/R$ . The observation of these HT–HP phases for the bulk material provides some clues as to the equilibrium structure of the cluster. The exact shape of the cluster will then depend on energy criteria and crystallographic factors.



**Fig. 7.1.** *Upper:* High resolution electron microscope image of a cubo-octahedral cobalt cluster with hexagonal faces, containing roughly 1 000 atoms [2]. *Lower:* Representation of the same polyhedron

### 7.1.2 Wulff Polyhedron

A cluster is made by assembling atoms, treated as spheres, with varying levels of order. This assembly of spheres cannot give rise to another sphere, but in fact produces a polyhedron that can be characterised by its outer faces. The shape and nature of the polyhedron depend on the binding energy of the atoms. In the case of a 3D crystallographic structure, the solution to this problem is given by the Wulff construction [1]. The construction criterion satisfies the following rule: if a face is characterised by Miller indices  $hkl$  and has area  $S$ , then

$$\frac{\gamma_{hkl}}{R_{hkl}} = \text{const.} \quad (7.10)$$

If there is no anisotropy, as in the drop model, where we have  $\gamma_{hkl}S_{hkl} = \gamma S$ , we simply obtain

$$\frac{\gamma}{R} = \text{const.} \quad (7.11)$$

This is the equation for a sphere (the droplet) because  $R$  must be constant. If the need for faces is taken into account, the construction becomes much more

difficult. It is nevertheless possible to estimate the shape of the polyhedron by means of a simple argument. Consider the face-centered cubic (fcc) structure. The coordination number of the lattice is 12, i.e., each atom has twelve neighbours. This is the most compact phase with the hexagonal close-packed (hcp) phase. Two polyhedra are associated with this structure: the Wigner–Seitz polyhedron (unit cell in the direct space) and its dual (Wigner–Seitz unit cell in the reciprocal space), known as the first Brillouin zone. The latter is a cubo-octahedron with hexagonal faces, known as a truncated octahedron, a semi-regular polyhedron with six square faces (100) and eight hexagonal faces (111) (see Fig. 7.1).

The (111) faces are the most densely packed with coordination number equal to nine. The (100) faces have coordination eight. Applying the Wulff criterion, for the two faces (111) and (100),

$$\frac{\gamma_{111}}{\gamma_{100}} = \frac{R_{111}}{R_{100}} . \quad (7.12)$$

In the bond cutoff model, characteristic of the Wulff polyhedron, the surface energy (in  $\text{J m}^{-2}$ ) is the product of a binding energy  $\xi$  and a density of broken bonds corresponding to a given face. The only important parameter is thus the number of atoms per unit area and the type of arrangement. According to (7.2), the area is a function of the number of atoms and depends only on their arrangement on the surface. Per unit area, the surface energy will increase with the number of atoms on this unit area. Introducing the coverage ratio  $o_c$  for the surface, the arrangement of atoms on a (111) face is of hexagonal close-packed type, so that  $o_c = \pi/\sqrt{12}$ , while for a (100) face it is of ‘square’ type, so that  $o_c = \pi/4$ . Hence,

$$\frac{\gamma_{111}}{\gamma_{100}} = \frac{o_{c(100)}}{o_{c(111)}} = \frac{\sqrt{3}}{2} = \frac{R_{111}}{R_{100}} . \quad (7.13)$$

Now, for a cubo-octahedron with hexagonal faces, the ratio of the distances between the square faces [(100) faces for the crystal] and the hexagonal faces [(111) faces for the crystal] is

$$\frac{R_{\text{hexagon}}}{R_{\text{square}}} = \frac{\sqrt{3}}{2} . \quad (7.14)$$

If the binding energy  $\xi$  does not vary much with the type of face, the Wulff polyhedron will not be far from the cubo-octahedron with hexagonal faces. Figure 7.1 shows an fcc cobalt cluster with this structure.

### 7.1.3 Beyond the Wulff Polyhedron

The energy of a cluster can be calculated by ab initio methods, or even by simple phenomenological methods. In order to approach this subject, let us consider the various types of chemical binding [3].

### Metallic Binding

The cohesive energy is provided by maximal overlap of atomic orbitals in order to delocalise the electrons. We outline the two most common models:

- The so-called bond cutoff model used in crystallography. In this model, the cohesive energy of an atom is proportional to the number of bonds and hence to the number  $C$  of nearest neighbours (the coordination number):

$$E_{\text{cohesive}} = CE_{\text{at}} , \quad (7.15)$$

where  $E_{\text{at}}$  is the binding energy of one atom per coordination. This intuitive model overestimates the energy.

- The tight binding model in the second moment approximation (TBSMA) [4]. The cohesive energy has the form

$$E_{\text{cohesive}} = \sqrt{C} \times E_{\text{at}} . \quad (7.16)$$

One expects the cohesive energy to be maximal when the surface area of the polyhedron is minimal and one expects the area of each face to be as densely packed as possible. This is the key property for understanding the equilibrium shape of small metallic clusters. It remains only to construct a polyhedron which satisfies this rule. However, there is no simple relationship between the two conditions. Let us examine each case.

#### *The Most Compact Polyhedron*



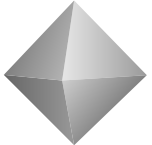
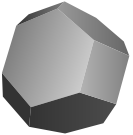
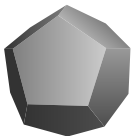

Just as in the approximation of the Wulff polyhedron, we start from the fcc structure. The compactness of a polyhedron can be calculated from the following criterion. Consider a set of points on a polyhedron occupied by some children. As the school is at the centre of the polyhedron, the problem is to minimise the mean squared distance that a child must cover to get to school, which is effectively the calculation of the second moment. Intuitively, it is clear that the sphere represents the perfect case. The minimum area criterion can be found in books on topology, tabulated for many different polyhedra. Figure 7.1 shows only the regular polyhedra and one semi-regular polyhedron, the Wulff polyhedron. The maximum compactness occurs for the icosahedron.

#### *Close-Packed Faces*

The dense packing of faces requires the point group of the polyhedron to be compatible with the space group of the lattice. Hence, the cube, octahedron, and truncated octahedron are compatible with the fcc structure. This raises the following question: is it possible to construct a unique polyhedron with the most densely packed (111) faces in our face-centered cubic lattice? The answer is affirmative: this construction leads to the non-truncated octahedron (see Table 7.1). In the case of the icosahedron, the point group  $I_h$  is not compatible. (Recall that the icosahedron has no translation symmetry and does not belong to the 230 groups of 3D systems.)



**Table 7.1.** Regular polyhedra and cubo-octahedron with hexagonal faces. The symmetry group and compactness are indicated. The second moment which characterises the compactness is normalised with respect to the sphere. A second moment close to unity implies a compact polyhedron. The formulas used to calculate the second moment can be found in [5]

Polyhedron	Shape	Symmetry	Compactness criterion	Type of face
Tetrahedron			1.351	(111)
Cube		m 3 m	1.083	(100)
Octahedron		m 3 m	1.073	(111)
Truncated octahedron		m 3 m	1.020	(111) and (100)
Dodecahedron		5 3 m	1.015	Incompatible
Icosahedron			1.011	Incompatible

### *Squaring the Circle*

Since the dense packing of a polyhedron and the corresponding faces are not directly correlated, a compromise has to be found. For this purpose, we must calculate (for example, using the bond cutoff model), the cohesive energy of

the cluster in order to find the most stable polyhedron. We shall consider just two polyhedra:

- The octahedron (see Table 7.1), which is not particularly compact but has close-packed (111) faces.
- The truncated octahedron (Wulff), which is more compact but has less densely packed (100) faces.

This analysis is then completed by considering special atoms with specific coordination numbers:

1. atoms belonging to edges,
2. atoms belonging to vertices.

These atoms play a fundamental role in catalytic processes. The coordination number of each atom can be found in the literature as a function of the number of atoms in a polyhedron ‘layer by layer’. This layer-by-layer quantification is clearly explained by the fact that a well-defined number of atoms is required to construct a homothetic polyhedron. Table 7.2 gives these values for the two polyhedra.

As in the liquid-drop model, we can introduce the dimensionless parameter  $E_{\text{norm}}$ . We have

$$E_{\text{cluster}} = \sum_i i \times E_{\text{at}} N_{(i)}, \quad (7.17)$$

where  $i$  is the coordination number. The cohesive energy per atom in the solid (coordination number 12) is

$$E_{\text{cohesive}} = 12 \times E_{\text{at}} N_{\text{T}}. \quad (7.18)$$

The energy per atom is then

$$E_{\text{norm}} = \sum_i i \times E_{\text{at}} N_{(i)} \times \frac{1}{12 \times E_{\text{at}} N_{\text{T}}} = \frac{\sum_i i \times N_{(i)}}{12 \times N_{\text{T}}}. \quad (7.19)$$

This can be compared with the value given in the liquid-drop model by (7.8). The results are shown in Fig. 7.2. The discrepancy is very small. For very small sizes, the truncated octahedron is the best candidate. From these considerations, we may deduce the following properties:

- The smaller the polyhedron, the lower the mean coordination number and hence the lower the cohesive energy.
- The energy difference is very low and there will be competition between different structures with the possibility of phase transitions.
- In nanoclusters, i.e., with fewer than 1 000 atoms, the contribution from special atoms located on edges and vertices is critical.

**Table 7.2.** Number of atoms with coordination number  $C$ , for  $C = 4, 7, 8, 9$  and  $12$  for two types of polyhedron as a function of the total number of atoms in the cluster. These values are taken from [6]

Octahedron				
Layer $m$	2	3	4	$m > 4$
Total number of atoms $N_T$	6	19	44	$m(2m^2 + 1)/3$
Number $N_{(4)}$ of atoms with coordination 4	6	6	6	6
Number $N_{(7)}$ of atoms with coordination 7	0	12	24	$12(m - 2)$
Number $N_{(9)}$ of atoms with coordination 9	0	0	8	$4(m - 3)(m - 2)$
Number $N_{(12)}$ of atoms with coordination 12	0	1	6	$(2m^3 - 12m^2 + 25m - 18)/3$
Cubo-octahedron with hexagonal faces				
Layer $m$	2	3		$m > 3$
Total number of atoms $N_T$	38	201		$16m^3 - 33m^2 + 24m - 6$
Number $N_{(6)}$ of atoms with coordination 6	24	24		24
Number $N_{(7)}$ of atoms with coordination 7	0	36		$36(m - 2)$
Number $N_{(8)}$ of atoms with coordination 8	0	6		$6(m - 2)^2$
Number $N_{(9)}$ of atoms with coordination 9	8	56		$8(3m^2 - 9m + 7)$
Number $N_{(12)}$ of atoms with coordination 12	6	79		$16m^3 - 63m^2 + 84m - 38$

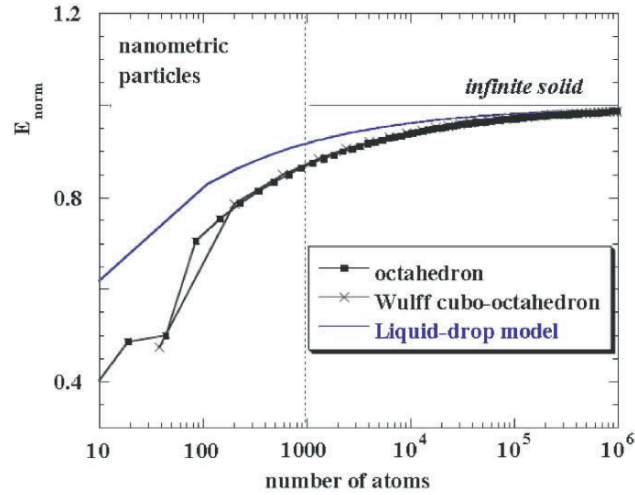
At this stage in the argument, the following question arises: is it possible to do better than the truncated octahedron?

The icosahedron, with the most compact shape (see Table 7.1), has two properties which a priori go against it:

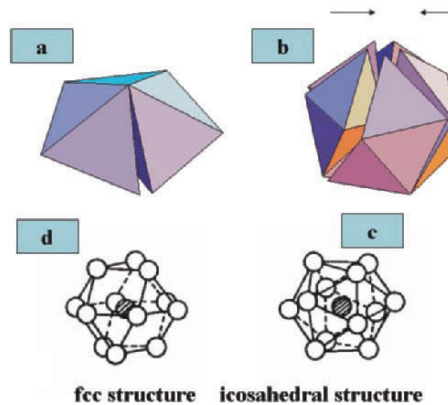
- it has no translation symmetry,
- the point group is  $I_h$ .

The first restriction is not fundamental for very small clusters, where the notion of periodicity has little meaning. The second can be negotiated as follows. There are three ways to generate an icosahedron:

- The first involves constructing an icosahedron by intersecting a family of planes of the type  $(hk0)$ , where  $h/k$  is close to the golden section  $\phi = 1/2 + \sqrt{5}/2$ , in an fcc lattice for example. Since the Miller indices are



**Fig. 7.2.**  $E_{\text{norm}}$  calculated using three different models. As the number of atoms increases,  $E_{\text{norm}}$  tends to unity, the value in the bulk. It is clear that the energy decreases very quickly when the size goes below 1000 atoms



**Fig. 7.3.** (a) Decahedron constructed from five tetrahedra. (b) Icosahedron constructed from twenty tetrahedra. (c) Smallest icosahedron with 13 atoms, compared with the fcc structure for the same cluster (d)

whole numbers, this is impossible. An approximate solution might be a family of planes (162 100 0)! This family of planes generates faces that are not densely packed and hence unfavourable as far as cohesion is concerned.

- The second solution involves constructing an icosahedron out of twenty judiciously stacked fcc tetrahedra (see Fig. 7.3). This provides a locally fcc structure with a twinned particle between each tetrahedron. This structure, known as a multitwinned particle (MTP) structure, is observed in small clusters. All the faces are close-packed (111) faces. However, an

absence of translation symmetry must arise here. Indeed, the stack of twenty tetrahedra cannot fill the whole of space, and an empty space known as the closing defect is left in the structure. This empty space is in fact filled by a more or less homogeneous relaxation of the atoms.

- The structure is a true icosahedron with symmetry  $I_h$  constructed around a central atom surrounded by two rings of five atoms (see Fig. 7.3). It can be shown that this icosahedron is obtained by deforming a cubo-octahedron with triangular faces. The closing defect is filled by introducing two distances, one radial and the other tangential. The separation is of the order of 5%. The reduction in overlap of the orbitals reduces cohesion but remains reasonable for very small dimensions.

The transition from a phase close to the Wulff polyhedron is a characteristic of nanostructures. Care must be taken not to confuse the decahedron and the icosahedron, both of which are multitwinned fcc structures with fivefold symmetry. They have been observed by diffraction and high resolution transmission electron microscopy (see Fig. 7.4). The decahedron observed along the axis of the fivefold symmetry clearly exhibits the five fcc sublattices. The icosahedron has a globular cluster structure which can only be identified by image simulation. Most of the observed structures correspond to decahedra. These structures are in principle less stable, but much easier to observe!

#### 7.1.4 Van der Waals Binding

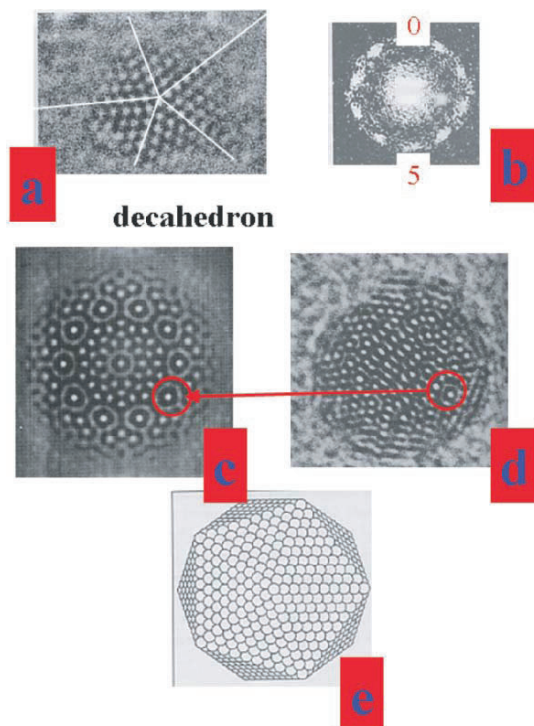
The van der Waals interaction is important mainly for the noble gases and assemblies of molecules in which the HOMO–LUMO gap is large, e.g., assemblies of fullerenes  $C_{60}$ . This is a very short range interaction. Using the classical approach of the two-body potential, we have in this case the standard Lennard-Jones potential  $E_{\text{SLJ}}$  given by

$$E_{\text{SLJ}} = E_{\text{cohesive}} = 2N\varepsilon \left[ \sum_j \left( \frac{\sigma}{p_{ij}R} \right)^{12} - \sum_j \left( \frac{\sigma}{p_{ij}R} \right)^6 \right], \quad (7.20)$$

where  $p_{ij}R$  is the distance from a given atom of the crystal labelled  $i$  to any other atom labelled  $j$  referred to the distance  $R$  between nearest neighbours. For the fcc structure, we find the values

$$\sum_j \left( \frac{1}{p_{ij}R} \right)^{12} = 12.132 = 20 \sum_j \left( \frac{1}{p_{ij}R} \right)^6 = 14.454. \quad (7.21)$$

The short range of this interaction validates the maximum compactness criteria discussed in the last section. The noble gases in the fcc bulk phase occur as icosahedra in small clusters. The icosahedral shape has in fact been observed in argon clusters.



**Fig. 7.4.** Second contrast image of a gold icosahedron (**d**) observed by very high resolution microscopy, compared with a decahedron (**a**) viewed along the axis of fivefold symmetry. (**b**) Micro-diffraction of the decahedron clearly revealing the fivefold symmetry (10 points). (**e**) Representation of the icosahedron. The typical globular structure of the icosahedron is difficult to image. An example is given in the figure. Images courtesy of M. Flueli [7]

### 7.1.5 Covalent Binding

Compactness is not a criterion for covalent bonding. Hence the diamond structure (Fd3m) has a compactness of 0.34, compared with 0.74 for the fcc structure. The optimisation criterion is geometric with the classical view of hybridisation. There are thus 1D linear structures (hybridisation  $sp$ ), plane structures (hybridisation  $sp^2$ ), and 3D structures (hybridisation  $sp^3$ ). These structures are characterised by the dihedral hybridisation angle (see Appendix B). Any angular separation introduces a destabilising elastic energy. Covalent structures are characterised by a gap between bonding and anti-bonding states, which varies from a fraction of an electronvolt to several electronvolts. This is relevant for structures of very small dimensions. The presence of atoms at the surface leads to unmatched bonds known as dangling bonds which introduce states into the gap. The existence of these states at the surface can have dramatic consequences as the size of the cluster decreases. To solve this

problem, the cluster will adopt a higher-dimensional geometry. The elastic energy expended to deform the lattice must be fully balanced by the suppression of surface states. We have the following three cases:

1. Hybridisation  $sp$ . The cluster adopts a circular shape in order to rid itself of all dangling bonds. This 2D structure is simply obtained by connecting together the atoms at the two ends of the initial linear structure.
2. Hybridisation  $sp^2$ . The cluster adopts a spherical shape. To do this, pentagons are introduced into the original hexagonal structure. These pentagons have the effect of curving the base plane. Euler's theorem [8] relates the number of vertices  $V$ , faces  $F$ , and edges  $E$  to obtain a convex polyhedron with all its vertices lying on a sphere:

$$V + F - E = 2 . \quad (7.22)$$

The hybridisation  $sp^2$  generates a hexagonal arrangement, which thus has a coordination number equal to 3. Pentagonal defects can be associated with it and these curve the hexagonal base plane into a convex lattice. One therefore seeks an association of  $p$  pentagons (with 5 edges) and  $h$  hexagons (with 6 edges), so that

$$F = p + h . \quad (7.23)$$

Each edge joins two vertices and each vertex shares an edge with three neighbours, whence

$$3V = 2E . \quad (7.24)$$

Each edge is common to two polygons, so that

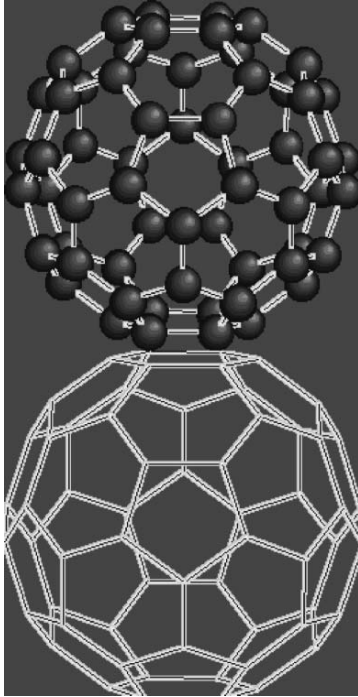
$$2E = 5p + 6h = 5p + 6(F - p) . \quad (7.25)$$

Hence, finally,

$$p = 12 . \quad (7.26)$$

This remarkable result is somewhat surprising. The number of pentagons must be exactly twelve, whilst the number of hexagons can be any even number. The best known example is fullerene  $C_{60}$ , which has twenty hexagons (see Fig. 7.5). The structure here is a truncated icosahedron in the shape of a football. The smallest is the pentagonal dodecahedron with no hexagonal faces, shown in Table 7.1. This is the basic structure of the clathrate compounds.

3. Hybridisation  $sp^3$ . As the basic shape is already 3D, the construction cannot increase the dimensionality of the lattice. One then observes a re-arrangement of the surface atoms in such a way as to minimise the states in the gap and increase the coordination number. This surface reconstruction can generate complex structures that can only be predicted by *ab initio* calculations. MTP structures have been observed for diamond, as in the case of metals. Filled fullerene-type structures are also envisaged.



**Fig. 7.5.** Representation of  $C_{60}$ : semi-regular polyhedron obtained by truncating an icosahedron by its dual, the dodecahedron

### 7.1.6 Ionic Binding

The ionic bond only exists in binary systems where a transfer of electrons between the two elements A and B is energetically favorable. The typical case is the association of two elements A and B belonging to columns I and VII of the periodic table, e.g., NaCl. The energy binding the atoms is the Madelung energy, whose main feature is that it is long range, being a Coulomb potential that goes as  $1/R$ . The total lattice energy of a crystal containing  $N$  anions and  $N$  cations is

$$E_{\text{tot}} = -\frac{N\alpha q^2}{R_{\text{eq}}} \left(1 - \frac{\rho}{R_{\text{eq}}}\right), \quad (7.27)$$

where  $q$  is the elementary charge,  $R_{\text{eq}}$  the distance between nearest neighbours,  $\alpha$  the Madelung constant found by summing over all attractive and repulsive interactions in the crystal, and  $\rho$  is the repulsion term, which is small compared with  $R_{\text{eq}}$ . There is no simple relation between the Madelung constant and the compactness ( $\alpha = 1.75$  for the NaCl structure and 1.76 for the CsCl structure). The key parameter is the relative size of the anions and cations and their arrangements in space. This explains why an ionic cluster conserves the structure it has in the bulk phase. For very small structures, where the numbers of cations and anions differ (odd number of atoms), a



defect forms which is analogous to the so-called centre defect  $F$  (missing negative ion with an extra electron bound to this hole).

## 7.2 Characteristic Quantity: Radius

An analysis of cluster properties taking full account of geometric features goes beyond the scope of this book. However, we can still give a fairly accurate description of what are known as size effects, forgetting the shape of the cluster and treating the problem with the liquid-drop model. The key parameter here is the radius  $R$  of the droplet (cluster). Any physical phenomenon is characterised by a coherence length  $\xi$ . If  $\xi$  is small compared with the cluster size as gauged by  $R$ , the observed phenomena will be the same as those observed in the bulk phase. Otherwise, if  $\xi \geq 2R$ , one expects the relevant physical quantity to depend on the size. This dependence can be analysed by noting that a cluster has two parts: a surface and a core. Since the ratio of the two is a function going as  $1/R$ , one expects the dependence of a physical quantity  $X$  for an object of radius  $R$  to have the form

$$X = \frac{X_{\text{bulk}}}{1 + \xi/2R} \approx X_{\text{bulk}} \left( 1 - \frac{\xi}{2R} \right). \quad (7.28)$$

We shall now illustrate this  $1/R$  dependence by several examples.

### 7.2.1 Thermodynamic Quantities: Melting Temperature

#### Liquid-Drop Model

The liquid-drop model considers a cluster containing  $N$  atoms. This can be in the solid phase (fixed atoms) or the liquid phase (atoms can move within the droplet). The solid–liquid transition involves among other things the disappearance of atomic planes and facets. In this sense, a liquid phase cluster will always be spherical. Gold clusters are liquid at room temperature despite the fact that the melting temperature of bulk solid gold is tabulated at 1336 K! We shall show that the notion of melting temperature depends on the cluster radius. In the simplest model, it is assumed that the cluster is spherical, homogeneous, and isotropic and comprises only one element. This cluster is characterised by its radius  $R$  alone. Let us express the chemical potentials as a function of the temperature and pressure. We shall use a first order expansion:

$$\mu(T, P) = \mu(T_0, P_0) + \frac{\delta\mu}{\delta T}(T - T_0) + \frac{\delta\mu}{\delta P}(P - P_0) + \dots \quad (7.29)$$

The thermodynamic quantities are related by the Gibbs–Duhem relation

$$-VdP + SdT + md\mu = 0, \quad (7.30)$$

where  $V$  is the volume,  $S$  the entropy, and  $m$  the mass. Hence,

$$\frac{\delta\mu}{\delta T} = -\frac{S}{m}, \quad (7.31)$$

$$\frac{\delta\mu}{\delta P} = \frac{V}{m} = \frac{1}{\rho}, \quad (7.32)$$

where  $\rho$  is the density. These quantities are valid for the liquid phase ( $P_1, T_1, \rho_1, S_1$ ) and the solid phase ( $P_s, T_s, \rho_s, S_s$ ). As  $P_0$  and  $T_0$  can be chosen arbitrarily, we shall consider the special case of the triple point in the bulk phase, for which we have the new equation

$$\mu_1(T_0, P_0) = \mu_s(T_0, P_0). \quad (7.33)$$

In this case,  $T_0$  is the melting temperature of the bulk phase. At the melting temperature  $T$  of the cluster, we have the equilibrium condition

$$\mu_1(T, P) = \mu_s(T, P). \quad (7.34)$$

From (7.29), one can calculate the chemical potential for the liquid and solid phases and express the difference between these two phases using (7.30)–(7.34) and introducing the term  $\Theta = T/T_0$  which expresses the discrepancy between the melting temperatures of the cluster and the bulk phase:

$$\begin{aligned} \mu_1(T, P) - \mu_s(T, P) &= \mu_1(T_0, P_0) - \mu_s(T_0, P_0) + \frac{\delta\mu_1}{\delta T}(T - T_0) \\ &\quad - \frac{\delta\mu_s}{\delta T}(T - T_0) + \frac{\delta\mu_1}{\delta P_1}(P_1 - P_0) - \frac{\delta\mu_s}{\delta P_s}(P_s - P_0). \end{aligned} \quad (7.35)$$

Introducing the latent heat of fusion  $L$  at constant pressure, viz.,

$$L = \frac{S_1 - S_s}{m} T_0, \quad (7.36)$$

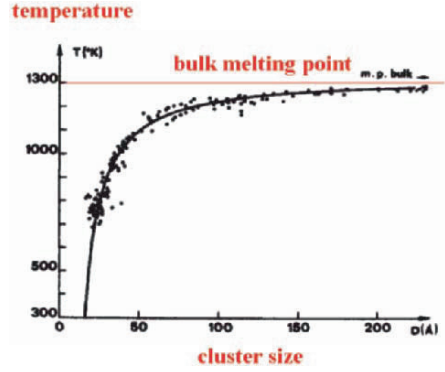
it follows that

$$0 = 0 + L(1 - \Theta) - \left( \frac{1}{\rho_1} - \frac{1}{\rho_s} \right) (P - P_0). \quad (7.37)$$

A pressure is exerted on the cluster known as the Gibbs pressure (derived from Laplace's law) [see (7.124)]. This pressure balances the force exerted on the curved solid. At the melting temperature,

$$P_1 = P_v + \frac{2\gamma_l}{R}, \quad (7.38)$$

where  $P_v$  is the saturated vapour pressure which can be neglected for high curvatures (small  $R$ ), such as one finds in clusters. Hence,



**Fig. 7.6.** Melting temperature of gold clusters as a function of cluster size: experiment and theory. The formula used is a second order expansion of (7.42) [9]

$$P_l - P_0 = \frac{2\gamma_l}{R_l}. \quad (7.39)$$

For the solid phase, an analogous relation can be deduced from the Wulff criterion (see Sect. 7.1.2):

$$P_s - P_0 = \frac{2\gamma_{s(hkl)}}{R_{(hkl)}}, \quad (7.40)$$

where  $R_{(hkl)}$  is the distance from a face of the polyhedron to the centre. In the liquid-drop model, we assume that  $R_{(hkl)}$  is equal to the mean radius in the solid phase, with the cluster taken as a sphere. In this case, there is a simple relation between  $R = R_l$  and  $R_s$ , viz.,

$$R_s = \left(\frac{\rho_l}{\rho_s}\right)^{1/3} R. \quad (7.41)$$

This leads to the final equation

$$1 - \Theta = \frac{2}{LR_s\rho_s} \left[ \gamma_s - \gamma_l \left(\frac{\rho_s}{\rho_l}\right)^{2/3} \right]. \quad (7.42)$$

This can also be written

$$T = T_m = T_0 \left( 1 - \frac{a}{R_s} \right), \quad (7.43)$$

where

$$a = \frac{2}{L\rho_s} \left[ \gamma_s - \gamma_l \left(\frac{\rho_s}{\rho_l}\right)^{2/3} \right].$$

**Table 7.3.** Numerical values of several parameters relevant to gold [9]

	Solid	Liquid
Density $\rho$ [ $\text{kg m}^{-3}$ ]	18 400	17 280
Surface tension $\gamma$ [ $\text{J m}^{-2} \text{K}^{-1}$ ]	1.38	1.135
Latent heat of fusion $L$ [ $\text{J kg}^{-1}$ ]		$6.27 \times 10^4$
Melting temperature $T_0$ [K]		1 336

For convenience, the radius  $R_s$  is taken in a first approximation as  $R$ . The dependence thus goes as  $1/R$ . The melting temperature decreases rapidly for clusters with diameters below 5 nm. The main difficulty here is to make accurate measurements of the surface tension in the solid and liquid phases. Table 7.3 gives the numerical values for gold.

According to this model, a cluster with radius 2 nm has a melting temperature of 880 K. The model can be improved by making a second order expansion of the chemical potential  $\mu(T, P)$ . This improved model is in good agreement with experiment (see Fig. 7.6) [9].

## 7.2.2 Electronic Quantities

### Ionisation Potential: Liquid-Drop Model

It is well known that an atom has a higher ionisation potential than the work function of a material made up of the same atoms. Assimilating it naively with a work function, one may ask how this ionisation potential varies with the cluster size as represented by  $R$ . We shall examine the case of metallic clusters which can be described using the liquid-drop model (see Appendix D). The ionisation potential can then be written

$$PI(R) = \Phi + \frac{3q^2}{32\pi\epsilon_0 R}, \quad (7.44)$$

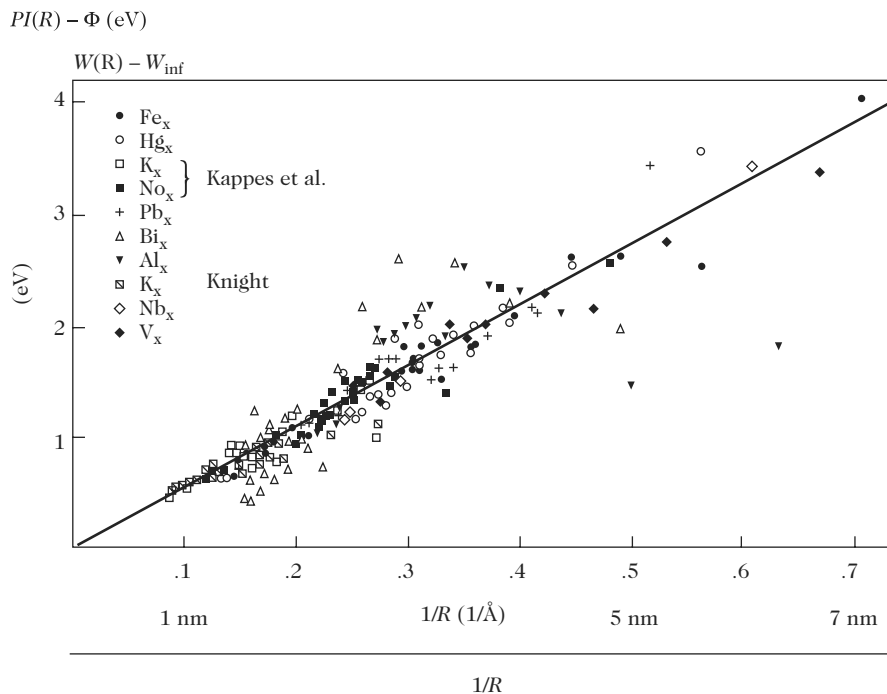
or alternatively,

$$PI(R) = \Phi + \frac{5.4}{R} \text{ eV}, \quad (7.45)$$

where  $R$  is given in  $\text{\AA}$ . The ionisation potential decreases as the cluster size increases. Figure 7.7 illustrates (7.45) for a selection of metals. This is partly due to charge screening, which is enhanced when the number of atoms in the cluster increases. The more the electron is screened, the less work needs to be done to remove it from the surface.

### Ionisation Potential: Hückel Model

Experiments carried out on a great many metallic clusters amply confirm this law. However, for very small sizes, one observes some distinct discrepancies



**Fig. 7.7.** Ionisation potential for a selection of metals as a function of cluster size: experiment and theory. The theoretical results are shown by the *continuous line*, which represents (7.45) [10]

from the monotonically increasing curve that corresponds to the  $1/R$  dependence. There are several reasons for these discrepancies:

- a geometric effect which is not taken into account by the liquid-drop model,
- a quantum effect due to the closing of electronic shells,
- an odd–even effect, perfectly described by the Hückel model, following from the fact that the stability of an even electron system is greater than the stability of a system with an odd number of electrons.

### Kubo Criterion

The metallic droplet model assumes that the cluster has the properties of the bulk material, i.e., electron delocalisation. In a metal, the electronic levels form a quasi-continuum which gives rise to a partially filled band. Levels tend to fill as far as the limit specified by the Fermi level  $E_F$ . In complete contrast, an atom or dimer is characterised by discrete atomic or molecular levels. One would intuitively expect the cluster to have intermediate properties, somewhere between the discrete and continuous systems. In other words,

one should expect a transition between these two states. By analogy with conduction properties, this is known as the metal–insulator transition [11].

To a first approximation, the mean spacing between two levels is inversely proportional to the number of atoms, i.e.,

$$\delta \approx E_F/N . \quad (7.46)$$

A transition should therefore be observed between the metallic-type quasi-continuum and the discrete system considered here as an insulator. In order to define a metallic-type state, an energy of at least  $\delta$  must be supplied so that the electron can occupy the empty upper state. This is possible if  $\delta$  is of the same order of magnitude as the thermal fluctuation  $k_B T$ , which gives the criterion

$$\delta = \delta_{\text{crit}} \approx E_F/N = k_B T . \quad (7.47)$$

More precisely, we may define  $\delta$  by

$$\delta = \delta E_F / \delta N = N(E_F)^{-1} = \delta_{\text{crit}} = k_B T , \quad (7.48)$$

where  $N(E_F)$  is the density of states at the Fermi level. This equation is valid if the cluster is neutral, whereupon

$$\delta_{\text{crit}} = k_B T < PI(R) , \quad (7.49)$$

or typically,  $T < 5 \times 10^5$  K. Moreover, the lifetime  $\tau$  of the electronic state must be long enough to define a level, i.e.,

$$\delta_{\text{crit}} \times \tau \geq \hbar , \quad (7.50)$$

implying a time  $\tau = 10^{-11}$  s for  $T > 0.8$  K. We shall now examine several specific cases.

#### *Monovalent Clusters (s Electrons)*

In this case, the free electron model is applicable. The Fermi level is given by

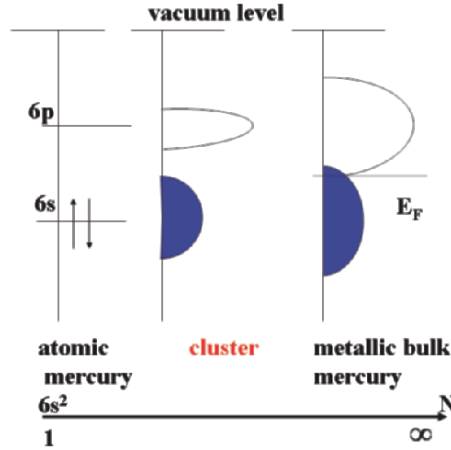
$$E_F = \frac{\hbar^2}{2m} 3\pi^2 n_e^{2/3} , \quad (7.51)$$

and the density of states at the Fermi level is

$$N(E_F) = E_F^{1/2} \frac{V}{2\pi^2} \frac{2m^{3/2}}{\hbar^2} , \quad (7.52)$$

where  $n_e = N/V$  is the electron density per unit volume. Letting  $d$  be the cluster diameter, we have the geometric relation

$$N = \frac{d^3}{(2r_a)^3} . \quad (7.53)$$



**Fig. 7.8.** Metal–insulator transition for mercury. In the case of a single atom (*left*), there is a  $6s$  state occupied by two electrons and an empty  $6p$  state. When the atoms cluster, the two discrete levels broaden to form two bands whose widths increase with the number of atoms  $N$ . In the solid, these bands are sufficiently broad to overlap and this then leads to conduction. If this overlap did not occur, the  $s$  band would be fully occupied and the  $p$  band would remain empty. We would then have a semiconductor. This is not the case since mercury is metallic. The transition occurs when the two bands come into contact

The transition occurs when

$$T = \frac{2\pi^2 \hbar^2 r_a^3 n_e^{2/3}}{m(3\pi^2)^{1/3} k_B R^3}, \quad (7.54)$$

where  $R$  is the radius in Å, whence

$$T = 0.564 \times 10^{-4} (n_e [\text{cm}^{-3}])^{2/3} \left(\frac{r_a}{R}\right)^3 [\text{K}]. \quad (7.55)$$

In the case of silver,  $n_e = 5.85 \times 10^{22} \text{ cm}^{-3}$ ,  $r_a = 3.02 \text{ Å}$ , which gives a critical radius of 1 nm if we wish to observe the metallic state at room temperature.

#### *Transition Metal Clusters (d Electrons)*

The same argument is valid in the case of the transition metals, provided we take into account the specific form of the  $d$  orbitals, which are much more localised in space, having a cigar shape, than the  $s$  electrons with their spherical distribution (see Appendix E).

#### *Divalent Metal Clusters (s and p Electrons)*

The divalent elements (Hg, Zn, etc.) have a full  $s$  shell. According to the free electron model, such elements must lead to completely filled bands and

hence to insulating structures. This is not actually the case, since the divalent elements are metallic in the bulk phase. The delocalisation of electrons arises from the overlap between the  $s$  and  $p$  bands, which grow broader as  $N$  increases, as happens for mercury. Figure 7.8 illustrates such a transition, which has a fundamentally different origin to the Kubo criterion. This kind of transition does not depend explicitly on the temperature.

#### *Covalent Clusters (s and p Electrons)*

In the case of elements whose  $s$  and  $p$  orbitals are hybridised, we obtain semiconductors in the bulk phase. This happens with the so-called  $sp^3$  structure, e.g., silicon. The Kubo criterion applies, even though there is no metal-insulator transition. As the size decreases, one simply observes an increase in the energy between the levels. This is the well known confinement model. Let us consider the evolution of the forbidden band or band gap  $E_{\text{tot}}$ . To keep things simple, we shall assume that the cluster is cubic with side  $L$ . It turns out that (see Appendix E)

$$E_{\text{tot}} \text{ [eV]} = E_g + \frac{0.75}{L^2 \text{ [nm]}}, \quad (7.56)$$

where  $E_g$  is the width of the band gap in the solid phase. The model breaks down as the cluster size decreases. For very small sizes, the  $sp^3$  hybridisation is no longer the stable phase because there are many dangling (unmatched) bonds at the cluster surface. If atomic states are accessible, e.g., the  $d$  states, hybridisation occurs between the  $s$ ,  $p$  and  $d$  states, with an increase in the mean coordination (the  $dsp^3$  hybridisation forms a tetragonal bipyramid). In this case, the forbidden band disappears and a transition occurs towards the metallic state. This phase transition is predicted for the elements in column IVA, with the exception of carbon, whose  $d$  states are inaccessible.

### 7.3 Characteristic Quantity: Fluctuations

In the last section we used several examples to illustrate the effect of the radius as a characteristic quantity. But there is another ‘hidden’ parameter which can lead to surprising effects in systems of very small size. This parameter is the fluctuation. Indeed, all ‘standard’ thermodynamics is based on the law of large numbers. The fact that a cluster contains a finite number of atoms renders certain theorems inoperable. We shall describe several examples to illustrate the consequences of these fluctuations, which are by far the most fascinating aspects encountered in the physics of nanosystems.

#### 7.3.1 Melting Temperature

The approach developed in the liquid-drop model is incomplete. Indeed, as we have seen above, the cohesive energy depends on the coordination number,



i.e., the number of nearest neighbours of an atom. The energy is lower at the surface than in the bulk. One therefore expects a lower melting temperature at the surface than in the bulk. This suggests introducing a two-phase system. In fact, the mechanism is much more complex than this. Close to the melting temperature, the cluster may very quickly change its equilibrium shape. Such fluctuations lead to atomic rearrangements involving displacements of the relevant atoms which may be likened to a liquid phase. There is no option but to go back to the basic principles of the thermodynamics of finite systems.

Consider a bulk system, i.e.,  $N \rightarrow \infty$ , in which the latent heat of fusion is  $L$ . Melting leads to an abrupt (first order) transition in the heat curve (internal energy as a function of temperature). This jump corresponds to  $L$ . Since the heat capacity  $C_v(T)$  is the derivative of the internal energy with respect to  $T$ , it will be a Dirac delta function  $\delta$  near  $T_0$ . We now ask what happens in a finite system. We shall make two hypotheses:

1. The system, comprising the  $N$  atoms of the cluster, is large enough to ensure that the entropy change is the same as in the solid. According to (7.36), we deduce that  $L$  is the same as in the solid phase.
2. The change in the melting temperature from  $T_0$  in the solid to  $T = T_m$  in the cluster [see (7.43)] only introduces a shift in the abscissa in the curve  $U(T)$ .

It can be shown that, for a finite system, a temperature fluctuation is observed such that

$$\langle \Delta T^2 \rangle = \frac{k_B T^2}{N C_v} . \quad (7.57)$$

In the thermodynamic limit ( $N \rightarrow \infty$ ), this fluctuation tends naturally to zero. Consequently, the phase transition occurs over a finite temperature range  $\Delta T$  in the vicinity of  $T_m$ . We may estimate  $\Delta T$  as follows. The specific heat is a Dirac delta function at the transition but keeps a finite value. The curve  $C_v(T = T_m)$  can be approximated by a box function such that

$$C_v(T) = \frac{\delta U}{\delta T} \sim \frac{L}{\Delta T_m} , \quad (7.58)$$

where  $L$  is the jump in  $U$  at  $T = T_m$ . Replacing  $C_v(T)$  by its value in (7.57), it follows that

$$\langle \Delta T^2 \rangle = \frac{k_B T_m^2 \Delta T_m}{N L} . \quad (7.59)$$

Hence, expressing  $L$  as a function of the entropy change given by (7.36),

$$L = \frac{S_l - S_s}{m} T_m = \frac{\Delta S}{m} T_m . \quad (7.60)$$

Assuming that the temperature fluctuation plays the intrinsic role of a temperature, which amounts to making the approximation

$$\langle \Delta T_m^2 \rangle = \langle \Delta T_m \rangle^2, \quad (7.61)$$

we obtain

$$\left\langle \frac{\Delta T_m}{T_m} \right\rangle = \frac{k_B}{N\Delta S} \propto N^{-1}. \quad (7.62)$$

The phase transition occurs over a temperature interval which varies as  $1/N$ . This result is a special case of an  $n$ th order phase transition, which for a finite system, leads to a variation that goes as  $N^{-\alpha}$ , where  $\alpha$  is an exponent which depends on  $n$ . The disappearance of the first order transition is characteristic of the thermodynamics of small clusters. A complete study of the transition can be made using Monte Carlo or molecular dynamics simulation techniques. The interval of melting temperatures corresponds to the coexistence of several solid and liquid phases. This effect has been demonstrated experimentally.

### 7.3.2 Kubo Model

The paramagnetic susceptibility of free electrons is independent of the temperature. This independence is directly related to the Fermi–Dirac distribution. In clusters, it has been shown that this susceptibility can vary with the temperature, depending on the valence of the element making up the cluster. These size effects follow directly from the pioneering work of Kubo on the statistical treatment of small numbers of electrons.

#### Paramagnetic Susceptibility

Let us examine the case of monovalent ( $s$  electron) elements at very low temperatures for which we therefore have  $\delta \gg k_B T$ . According to the Kubo criterion (7.48), the metal is therefore in an insulating phase. If we consider a very low temperature, only the ground state (the Fermi level) and the first few states above it can be occupied. Because there is a finite number of atoms, we must introduce the idea of a fluctuation in the energy levels. This is particularly true of the surface atoms, whose energy spectrum may exhibit fluctuations with respect to atoms in the bulk. There is a statistical distribution of levels around an energy level  $\varepsilon$  close to the Fermi level. The separation between two levels is a sequence  $\Delta_n$ ,  $n = \dots, -2, -1, 0, 1, 2, \dots$ , where  $\Delta_n \ll \delta$ . If the levels are equidistant to a first approximation, i.e.,  $\Delta_n = \Delta$ , we have a random distribution of energy levels given by the Poisson law

$$P_0(\Delta) = \frac{\exp(-\Delta/\delta)}{\delta}. \quad (7.63)$$

For a large system ( $N \rightarrow \infty$ ), the fluctuation term represented by  $\Delta$  grows smaller, i.e.,  $\Delta/\delta \rightarrow 0$ , whence it follows that

$$P_0(\Delta) = \frac{1}{\delta} = N(E_F), \quad (7.64)$$

according to (7.48). We thus recover the classical distribution of free electrons in an infinite solid.

Let us now ask what happens when a magnetic field  $\mathbf{H}$  is applied. In this case, level degeneracy is removed by the Zeeman effect. The separation between levels occupied by electrons with spins  $S = 1/2$  and  $S = -1/2$  is  $g\mu_B H$ , where  $\mu_B$  is the Bohr magneton and  $g$  the Landé factor ( $g = 2$  for an  $s$  state). The magnetic field is assumed weak enough to satisfy

$$2\mu_B H \ll \Delta. \quad (7.65)$$

Since the number of electrons is finite, we shall use the canonical ensemble to define the electron magnetic susceptibility  $\chi_e$ . By definition, this ensemble is given in terms of the partition function  $Z$  by

$$\chi_e = \lim_{H \rightarrow 0} k_B T \frac{\delta^2}{\delta H^2} \langle \ln Z \rangle, \quad (7.66)$$

where the partition function is summed over the states  $i$ ,

$$Z = \sum_i \exp(-E_i/k_B T). \quad (7.67)$$

In principle, this sum is taken over an infinite number of values of  $i$ . Since we consider only the first excited states, the sum goes to the last level which may be occupied, labelled  $i_{\max}$ . In practice, a good approximation is obtained with  $i_{\max} = 5$ .

There are two cases depending on the parity of the cluster.

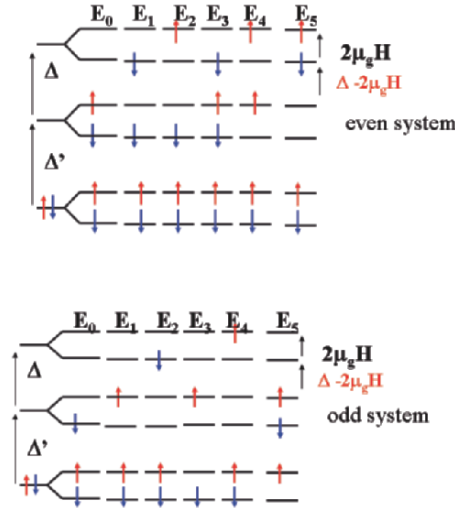
#### *Divalent Elements: Magnesium*

The atomic structure is a closed shell with valence  $s^2$  and therefore an even number of electrons. We shall consider only the two valence electrons (upper part of Fig.7.9). The ground state is specified by the energy  $E_0$ . The five excited states have energies

$$\begin{aligned} E_1 &= E_0 + \Delta - 2\mu_B H, & E_2 &= E_3 = E_0 + \Delta, \\ E_4 &= E_0 + \Delta + 2\mu_B H, & E_5 &= E_0 + 2\Delta. \end{aligned} \quad (7.68)$$

According to (7.67) and (7.68),

$$\begin{aligned} Z_{\text{even}} &= Z \\ &= \exp\left(-\frac{E_0}{k_B T}\right) \left[ 1 + 2 \exp\left(-\frac{\Delta}{k_B T}\right) \left( 1 + \cosh \frac{2\mu_B H}{k_B T} \right) \right. \\ &\quad \left. + \exp\left(-\frac{2\Delta}{k_B T}\right) \right]. \end{aligned} \quad (7.69)$$



**Fig. 7.9.** Ground state and first excited states in clusters with an even number (*upper*) and an odd number (*lower*) of electrons. The corresponding energies are indicated

It follows that (see Appendix F)

$$\chi_{\text{even}} = 3.04\mu_B^2/\delta = N(E_F)3.04\mu_B^2 . \quad (7.70)$$

In this model, the susceptibility is independent of the temperature. Note the similarity with the Pauli magnetic susceptibility for free electrons, viz.,

$$\chi_{\text{Pauli}} = \mu_B^2 N(E_F) . \quad (7.71)$$

However, experiment shows a temperature dependence. This dependence arises as soon as we assume that the levels are not equidistant, so that they do not have a Poisson distribution of the type given by (7.64), but a distribution of type

$$P_1(\delta) = P_0(\delta) \frac{\Delta}{\delta} , \quad (7.72)$$

for example. This is the so-called orthogonal distribution. In this case, numerical integration gives the susceptibility as

$$\chi_{\text{even}} = 7.63\mu_B^2 k_B T / \delta^2 . \quad (7.73)$$

Other mathematical distributions  $P_2(\delta)$ ,  $P_3(\delta)$ ,  $\dots$ ,  $P_n(\delta)$  are also used in the literature. The idea is to obtain better and better approximations to the true distribution of the levels  $\Delta$  in order to refine the model. However, the basic idea remains the same as the one we have described above.

*Monovalent Elements: Silver*

The argument is the same here. Consider the ground state and first excited states (see Fig. 7.9). The ground state has energy  $E_0 - \mu_B H$ . Using the general model in which the energy levels are not equidistant, the first excited states are

$$\begin{aligned} E_1 &= E_0 + \mu_B H , \\ E_2 &= E_0 + \Delta - \mu_B H , \\ E_3 &= E_0 + \Delta_{-1} - \mu_B H , \\ E_4 &= E_0 + \Delta + 2\mu_B H , \\ E_5 &= E_0 + \Delta_{-1} + 2\mu_B H . \end{aligned} \quad (7.74)$$

The partition function is

$$\begin{aligned} Z_{\text{odd}} &= Z \\ &= 2 \exp\left(-\frac{E_0}{k_B T}\right) \cosh\left(\frac{\mu_B H}{k_B T}\right) \left[1 + \exp\left(-\frac{\Delta}{k_B T}\right) + \exp\left(-\frac{\Delta_{-1}}{k_B T}\right)\right] . \end{aligned} \quad (7.75)$$

In the same way, we obtain

$$\chi_{\text{odd}} = \frac{\mu_B^2}{k_B T} . \quad (7.76)$$

In monovalent systems, the susceptibility is the Curie susceptibility which varies as  $1/T$ . This dependence has been confirmed experimentally.

**Heat Capacity**

The same model can be used to calculate the heat capacity using the definition

$$C = k_B \beta^2 \frac{\delta^2 \ln Z}{\delta \beta^2} , \quad (7.77)$$

where  $\beta = 1/k_B T$ . Taking the Poisson distribution  $P_0(\Delta)$ , we obtain

$$C_{\text{even}} = 5.02 k_B^2 T / \delta = 5.02 k_B^2 T N(E_F) , \quad (7.78)$$

$$C_{\text{odd}} = 3.29 k_B^2 T / \delta = 3.29 k_B^2 T N(E_F) . \quad (7.79)$$

This should be compared with the heat capacity of an electron gas (infinite solid) given by

$$C_{\text{electron}} = \frac{1}{3} \pi^2 k_B^2 T N(E_F) = 3.29 k_B^2 T N(E_F) . \quad (7.80)$$

For a monovalent element, we recover the electron heat capacity.

To sum up, divalent and monovalent systems behave in very different ways. This illustrates the effect of electron correlations in very small systems. The same phenomenon is observed for the odd–even effect in the ionisation potential (see Sect. 7.2.2).

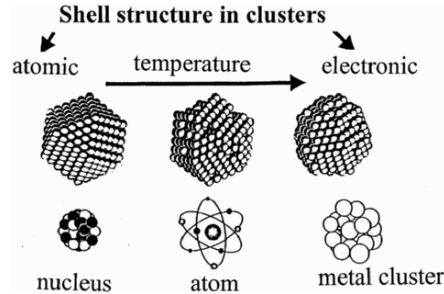
## 7.4 Specific Quantum Effects in Nanoscale Systems and Collective Excitations

In this section we discuss quantum effects related to the finite size of clusters, as observed in the properties of metallic clusters. These effects result directly from the quantisation of the electronic structure [12–16]. They concern the electronic and optical properties, which are directly correlated with the electronic structure in nano-objects. The common theme in the various parts of this section is to understand the way physical properties evolve with size, from the atomic scale to a scale of a few nanometers. Beyond 10 nm (as a rough order of magnitude, depending on the property under consideration), the nano-object is large enough to justify a classical approach, e.g., the theory of continuous media, classical macroscopic description of polarisation, etc., or a quantum description appealing to condensed matter physics, e.g., electronic band structure.

Section 7.4.1 deals with electron shell structure in clusters. There is a parallel here with the analogous structures observed in atoms or atomic nuclei (nucleons rather than electrons in the last case). Most of the important ideas will in fact be introduced in this part. Section 7.4.2 deals with electronic supershell structure as observed in the larger size range, i.e., more than about 1 000 atoms. Shell and supershell structures lead to non-monotonic evolution of physical properties when the size of the nano-object increases. Section 7.4.3 examines the specific optical properties of small metallic clusters in contrast to those of the bulk solid (dielectric confinement effect), and the associated quantum size effects.

Apart from the geometric structure due to atomic stacking discussed in Sect. 7.1, the physics of metallic clusters  $M_N$  is governed by their electronic properties. As the electronic and geometric structures are often closely inter-related, attempts to correlate the evolution of a physical property with one or other of these two structures are often dictated by criteria of convenience and simplicity.

In nano-objects, the surface to volume ratio is high. Indeed, about half of the atoms are located on the surface for  $N \sim 1\,000$ . Adjoining one or a few atoms to the surface of a small cluster may significantly alter its global symmetry, its mean coordination number and hence also its ionisation potential (IP), its stability with regard to fragmentation, and so on. Experimentally, depending on the element, the size range, and fabrication conditions (especially the temperature), the non-monotonic evolution of physical properties can often be directly correlated with the steady growth of a faceted structure with the shape of a regular polyhedron such as the Wulff polyhedron, icosahedron, cubo-octahedron, etc. (see Sect. 7.1). In contrast, for very small clusters, liquid clusters or clusters with disordered structure, and more particularly the ‘simple’ metal clusters (alkali metals, noble metals, and certain divalent metals in column IIB), the observed properties and their evolution with size depend little on the granular structure, but directly reflect the quantisation of



**Fig. 7.10.** Universality of fermionic shell structure in systems made up of fermions (see Appendix G), i.e., electrons in the case of atoms or metallic clusters, and nucleons in the case of atomic nuclei. This structure arises from the quantisation of energy levels when the system is spatially confined. In metallic clusters, the temperature often arbitrates the competition between atomic and electronic shell structure

electronic states confined in the interior of the cluster. A description of type ‘quantum liquid drop’ is then appropriate, treating the system as a positively charged ion distribution (mean charge density  $n_0$ ) interacting with a gas of ‘quantum’ electrons delocalised over the finite volume of the cluster (conduction electrons of the metal). This simplified physical picture suffices to explain all the ideas in this section.

#### 7.4.1 Electronic Shell Structure

The appearance of fermionic shell structure in systems made up of fermions (see Appendix G) is a general feature in physics (see Fig. 7.10), both on the macroscopic scale (electronic band structure which leads directly to the classification of different materials into conductors, insulators and semiconductors), and on the microscopic scale, i.e., the atom (structure which led to the classification of the elements into the periodic table at the end of the nineteenth century, well before the advent of the quantum theory), and the atomic nucleus (magic numbers corresponding to highly stable nuclei, see Appendix G).

As in atomic or nuclear physics, the shell structure in clusters is a consequence of the quantisation of electronic (or nucleonic) levels in a spherically symmetric effective potential known as the mean field. The analogy with nuclear physics is quite remarkable insofar as calculations show that the form of this potential follows, to a first approximation, the Woods–Saxon curve:

$$V_{\text{eff}}(r) = -\frac{V_0}{1 + \exp\left[(r - R_{N_e})/a\right]}, \quad (7.81)$$

where  $V_{\text{eff}}$  is a potential well of finite depth  $V_0$  and with a flat bottom which is rounded at the edges in a way characterised by the skin parameter  $a$ , similar to the potential well confining nucleons in the nucleus (see Fig. 7.11). The

quantity  $R_{N_e} = r_s N_e^{1/3}$  is the cluster radius,  $N_e$  is the number of confined conduction electrons, and  $r_s$  is the Wigner–Seitz electron radius, a parameter characterising the mean volume per conduction electron for the relevant element. This radius is related to the mean charge density per unit volume by

$$n_0 = q \frac{3}{4\pi r_s^3}. \quad (7.82)$$

In order to simplify the following discussion, it will be useful to begin by presenting the theoretical predictions made in the framework of the self-consistent jellium model. This model provides an excellent approximation with which to describe the electronic shell structure in clusters of simple metals, at least on a qualitative level. The discrete ionic structure is modelled by a homogeneous, spherically symmetric, positive charge density with steep sides (see Fig. 7.12).

The charge density is then given by

$$n_+(r) = n_0 H(R_{N_e} - r), \quad (7.83)$$

where  $H(x)$  is the step function defined by  $H(x > 0) = 1$  and  $H(x < 0) = 0$ , and  $n_0$  is the mean charge density of ions in the bulk phase of the element. The electronic structure of the ground state is generally calculated using the DFT formalism (density functional theory), by solving the Kohn–Sham (KS) equations self-consistently (see Chap. 18 for more details). The term ‘self-consistent’ means that the effective potential well  $V_{\text{eff}}(r)$  which confines the electrons spatially is not imposed but depends on the solutions of the Kohn–Sham equations through the electron density  $n(r)$ . In this way, we find

$$\left[ \frac{-\hbar^2}{2m} \nabla^2 + V_{\text{eff}}(r, n) \right] \Psi_i(\mathbf{r}) = \varepsilon_i \Psi_i(\mathbf{r}), \quad (7.84)$$

where

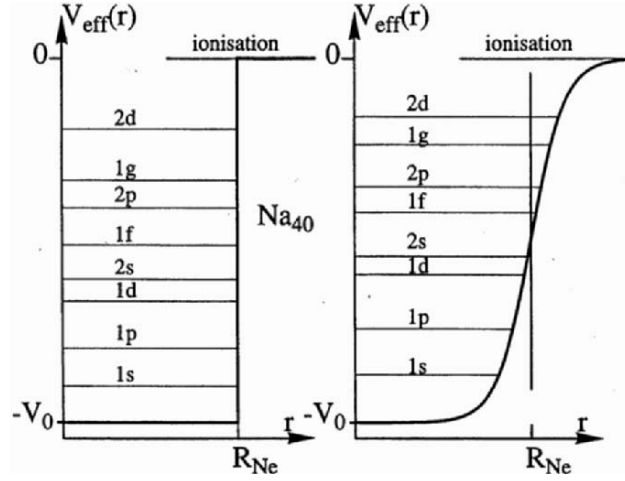
$$n(r) = -q \sum_{i=1}^{N_e} |\Psi_i(\mathbf{r})|^2$$

and

$$V_{\text{eff}}(n, r) = V_{\text{jel}}(r) + V_{e-e}(r).$$

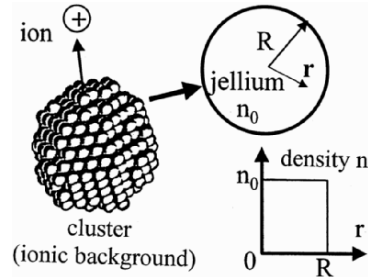
The quantity  $n(r)$  is the (negative) electron charge density expressed in terms of the squared amplitude of the occupied Kohn–Sham orbital states  $\Psi_i(\mathbf{r})$ . The latter are given as the product of a radial function  $R_{n,l}(r)$  and a spherical harmonic  $Y_l^m(\theta, \varphi)$ . The function  $V_{\text{eff}}(r)$  is the effective electronic confinement potential, the sum of the interaction  $V_{\text{jel}}(r)$  between the electron and the positively charged jellium ( $Q$  is the total positive charge) and a purely electronic contribution  $V_{e-e}(r)$  which includes the exchange correlation term, apart from the classical Coulombic interaction between the electron and the other electrons.  $V_{\text{jel}}(r)$  is given by





**Fig. 7.11.** Examples of central potential wells  $V_{\text{eff}}(r)$  with simple shapes, used to provide an approximate description of the electronic shell structure in metallic clusters, illustrated here for the sodium cluster  $\text{Na}_{40}$  comprising  $N = 40$  atoms, which therefore possesses  $N_e = 40$  delocalised conduction electrons. *Left:* Square potential with finite depth, flat bottom ( $-V_0$ ), and vertical walls. *Right:* potential with finite depth, flat bottom ( $-V_0$ ), and rounded, slightly sloping walls. In the second case, the shape is of Woods-Saxon type given by  $V_{\text{eff}}(r) = -V_0/\{1 + \exp[(r - R_{N_e})/a]\}$ , where  $r$  is the distance of the electron from the cluster centre and  $R_{N_e}$  is the cluster radius. The steepness of the sides of the potential well is characterised by  $a$ . In a central potential, the electron energy levels are labelled by two quantum numbers.  $n$  is the principal quantum number, a positive integer, and  $l$  is the azimuthal quantum number, associated with the orbital angular momentum of the electron:  $l = 0$  (denoted by  $s$ ),  $l = 1$  (denoted by  $p$ ),  $l = 2$  (denoted by  $d$ ),  $l = 3$  (denoted by  $f$ ),  $l = 4$  (denoted by  $g$ ), and so on. For example, the level  $1p$  is the lowest level with angular momentum  $l = 1$ . The levels have degeneracy given by  $g = 2(2l + 1)$ . The ground electronic state of the cluster is obtained by putting all  $N_e$  electrons in the  $N_e$  lowest quantum states. For the cluster  $\text{Na}_{40}$ , the electronic configuration is therefore  $(1s^2)(1p^6)(1d^{10})(2s^2)(1f^{14})(2p^6)$ , where the superscript indicates the number of electrons occupying the level. Note that the position of the levels depends on the profile of the effective potential  $V_{\text{eff}}(r)$ . For large sizes  $N$ , the order also depends critically on this feature. Hence, in order to calculate the electronic shell structure, one must determine, for each size, the corresponding self-consistent effective potential, like those in Fig. 7.13. The term ‘self-consistent’ implies that the shape of the potential is not imposed a priori, but depends on the electronic configuration through the associated electron density

$$V_{\text{jel}}(r) = \begin{cases} \frac{-qQ}{4\pi\epsilon_0 r} & \text{for } r \geq R_{N_e} , \\ \frac{qQ[(r/R_{N_e})^2 - 3]}{8\pi\epsilon_0 R_{N_e}} & \text{for } r \leq R_{N_e} . \end{cases} \quad (7.85)$$



**Fig. 7.12.** Jellium model. The metal cluster  $M_N$  is treated as a globally spherical pile of  $N$  positive ions and a gas of  $N_e$  delocalised conduction electrons. The discrete ionic structure cannot be taken into account in calculations, except for very small clusters. The jellium model consists in replacing the granular ionic structure by a spherical, homogeneous, positive charge distribution  $n_+(r)$  (charge density  $n_0$  for  $r < R$ ).  $n_0 = Q/(4\pi R^3/3)$ , where  $R$  is the radius of the sphere and  $Q$  the total positive charge.  $Q = Nwq = N_e q$ , where  $w$  is the valence of the element, i.e., the number of conduction electrons provided by each atom, e.g.,  $w = 1$  for the alkali metals,  $w = 3$  for aluminium and gallium

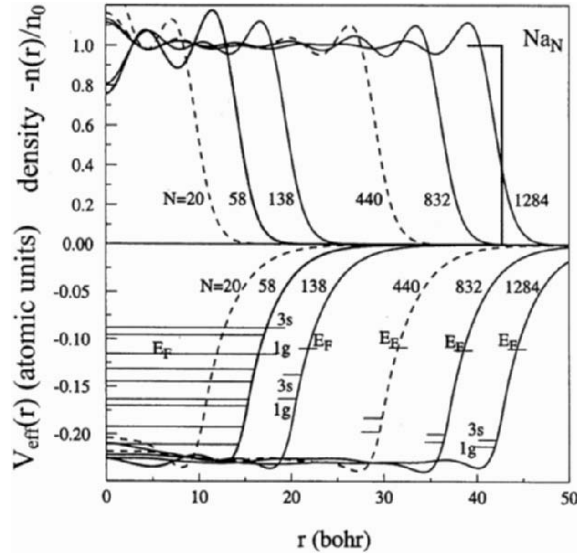
When convergence is reached, the system can be considered formally as equivalent to a system of  $N_e$  independent electrons moving in a central potential  $V_{\text{eff}}(r)$ . This simplified physical picture suffices to justify all the following discussion from an intuitive standpoint. Figure 7.13 gives the relevant results obtained in the framework of this model (single parameter  $r_s$ ), for neutral clusters of sodium  $\text{Na}_N$  ( $r_s = 3.93$  bohr).

The shell structure, which can be observed directly with a mass spectrometer (analysing the relative abundance of the different sizes), results from:

- the high degeneracy  $g = 2(2l + 1)$  of the electronic levels  $\varepsilon_i = E_{n,l}$ , where  $l$  is the angular momentum,
- the fact that the profile of the effective potential is almost independent of the size.

Sizes with relatively high stability, referred to as magic sizes by analogy with what happens in nuclear physics, are those for which all levels are completely filled (closed), i.e., there are  $2(2l + 1)$  electrons in the level. These sizes are produced in greater abundance by cluster sources (see Fig. 7.14).

The second condition here, which leads to the emergence of the shell structure clearly illustrated in Fig. 7.13, is essential for the approximate preservation of the order of the  $E_{n,l}$  levels and therefore ensuring that they fill up sequentially as the cluster size increases. (The electron ground state is obtained by filling the  $N_e$  lowest individual quantum states.) In fact, calculation shows that the order of the levels is not universal, except for very small sizes where the gaps between successive levels are large, but is modified for energies corresponding to high densities of states. This reorganisation from one size to the next is a consequence of the finite depth  $V_0$  of the effective potential, but



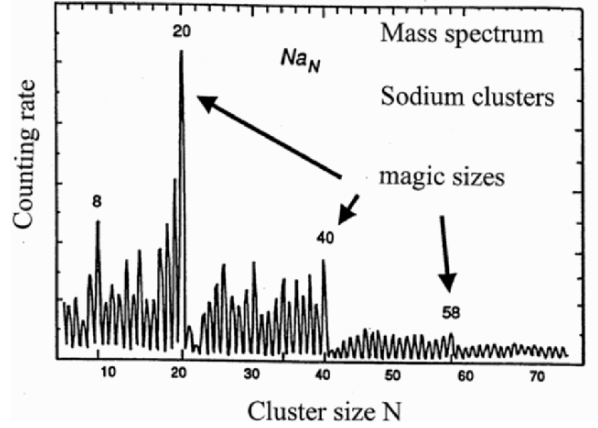
**Fig. 7.13.** Electron densities  $n(r)$  and self-consistent effective potentials  $V_{\text{eff}}(r)$  calculated using the jellium model (see Fig. 7.12) for magic clusters of sodium ( $N_e = N = 20, 58, 138, \text{etc.}$ ). The occupied electron levels and also the first two unoccupied levels of the  $\text{Na}_{58}$  cluster are indicated by *horizontal lines*. The dependence of the Fermi level  $E_F$  (last occupied level), the  $1g$  level (Fermi level of  $\text{Na}_{58}$ ) and the  $3s$  level on the size of the cluster are shown by *short dashes*. This reveals a steady decrease in the levels toward the bottom of the potential well, which can be modelled by the approximate scaling law  $E + V_0 \sim 1/R_{N_e}^2$ , where  $R_{N_e}$  is the radius of the jellium. The levels also tend to condense since the gaps between them narrow with increasing size. The edge of the jellium is shown for the  $\text{Na}_{1284}$  cluster. Note that the shapes of the curves  $n(r)$  and  $V_{\text{eff}}(r)$ , and also the bottom  $-V_0$  of the effective potential and the Fermi level are largely independent of the size

also of its non-step-like profile as modelled by the Woods–Saxon potential (see Figs. 7.11 and 7.13). This effect can be compared with the not totally sequential filling of the  $nd$  and  $(n+1)s$  electronic shells from one atom to the next on moving through the transition elements.

This reorganisation phenomenon with regard to the order of the levels, which is significant in clusters, does not fundamentally destroy the shell structure. However, it modifies the sequence of highly stable magic clusters calculated using a model in which the well is step-like and infinitely deep (see the left-hand diagram of Fig. 7.11, with  $V_0$  infinite), in which case the levels, measured from the bottom of the well, follow a perfect scaling law of the type

$$E_{n,l} = \frac{a_{n,l}}{R_{N_e}^2}. \quad (7.86)$$

In the very small size range, the almost universal order of the levels is



**Fig. 7.14.** Mass spectrum of hot sodium clusters produced by a heat source, revealing the electronic shell structure [17]. The magic numbers  $N$  (sizes involving closed  $E_{n,1}$  levels) are followed by a sudden decrease in peak intensities, the intensity being proportional to the number of clusters having a given size. The size  $N + 1$  immediately above a magic number corresponds to the beginning of the filling of an electron level that is less deep in the potential well (see Figs. 7.11 or 7.13 as an intuitive guide). The presence of this less strongly bound electron has the effect of reducing the stability of the  $N + 1$  cluster, as compared with the magic cluster  $N$ . One or more atoms will then evaporate from the  $N + 1$  cluster and size  $N + 1$  will therefore be detected with low abundance. The figure has been constructed using the experimental spectrum in [18]

$$1s/1p/1d/2s/1f/2p/1g/\dots,$$

leading to the magic sizes

$$N_e = 2, 8, 18, 20, 34, 40, 58, \dots$$

Note that the magic nature of a cluster is modulated by the proximity and clumping of successive levels, characteristic of this type of flat-bottomed potential. In practice, the observed magic numbers reveal the presence of a large gap between the last occupied electronic level (the Fermi level  $E_F$ ), which is completely filled (closed), and the first unoccupied level. For example, the  $1d$  and  $2s$  levels, which have similar energies (see Fig. 7.13), behave like a single level with degeneracy  $g = 12$ , and the magic number  $N_e = 18$  is rarely observed.

By considering Fig. 7.13, a parallel can be established between:

- the ‘magic’ numbers  $N_e$  and rare gas atoms, characterised by an increased relative stability (all electrons strongly bound) and a high ionisation potential, etc.,
- the numbers  $N_e + 1$  and the alkali atoms (presence of a weakly bound electron, high chemical reactivity, low ionisation potential), which corresponds to the filling of a new electron shell with higher energy.

Because of these striking analogies, the metallic clusters are sometimes referred to as giant atoms. In a quite general way, the dependence of the physical properties of clusters, such as the ionisation potential, electron affinity, or binding energy per atom, on the cluster size can be parametrised in the form of a sum of a slowly varying expansion of the type given by the classical liquid-drop model (sum of a bulk term, a surface term and possibly a curvature term, etc.), dominating at large sizes, and a fluctuating term reflecting the quantised electron structure, which leads to the specific properties of very small clusters. More succinctly, if  $A$  denotes such a physical property, the generic dependence on size takes the form

$$A(N) = a_B + \frac{a_S}{N^{1/3}} + \frac{a_C}{N^{2/3}} + \Delta A_{\text{shell}}(N). \quad (7.87)$$

This is a generalisation of (7.28). As for atomic nuclei, the mean dependence, restricted here to the dominating surface and curvature terms, is a consequence of the simple scaling law followed by the density and effective potential (see Fig. 7.13). In particular, for the ionisation potential, the asymptotic value  $a_B$  corresponds to the work function of the bulk metal and the surface term  $a_S/N^{1/3}$  is very close to the classically calculated energy required to move to infinity an electron initially located in the vicinity of the metal surface  $[(3/8)(q^2/4\pi\epsilon_0 R_{N_e})]$ .

The first experimental confirmation of the development of an electronic shell structure in metallic clusters was obtained with the alkali elements. Intense beams of clusters of these elements could be produced due to their low melting point, using simple heat sources. Moreover, their electronic structures in the atom and the bulk phase are particularly simple, considerably simplifying the theoretical analysis. The highly stable magic numbers  $N_e$  (equal to  $N$  for the monovalent element sodium), immediately followed by a sudden drop in the signal (the extra electron has to go into the next shell, with much higher energy, whereupon it is much more weakly bound) are clearly apparent in the mass spectrum of Fig. 7.14. The shell structure was then observed in the monovalent noble elements [electronic structure in the atom  $nd^{10}(n+1)s$ , Au, Ag], the divalent elements of group IIB [ $nd^{10}(n+1)s^2$ , Cd, Hg], and the trivalent elements of group IIIA [ $ns^2p^1$ , Ga, In].

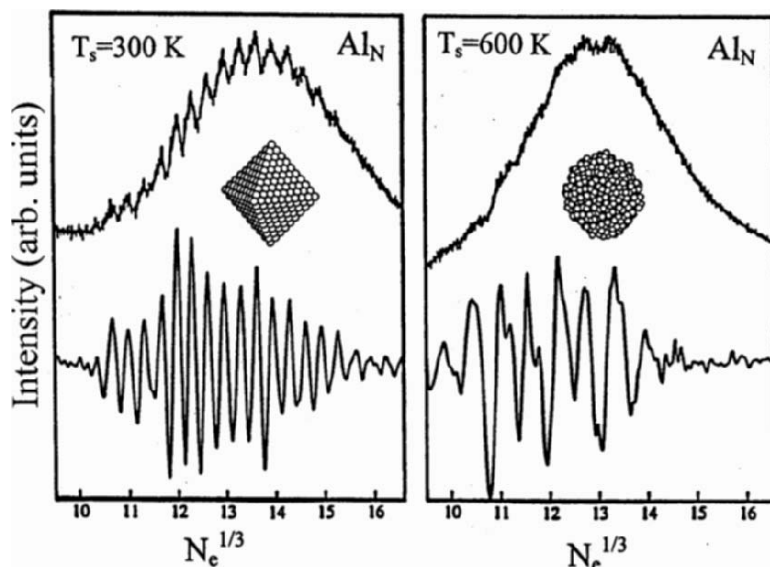
Whether it be in the unprocessed mass spectra or in the size dependence of a physical property [contribution  $\Delta A_{\text{shell}}(N)$  in (7.87)], the shell structure generally shows up either through striking saw-tooth ‘anomalies’ in the small size range or low amplitude oscillations in the large size range. The loss of contrast for large clusters is expected due to the clumping of electronic levels in the vicinity of the Fermi level (narrowing of the gaps between successive levels). Temperature effects leading to an effective broadening of electron levels, shape fluctuations removing the degeneracy of the  $E_{n,l}$  levels, populated rotational–vibrational levels, together with the huge number of isomers of the same energy which are probably produced in the cluster sources, are other factors leading to loss of contrast.

Specific experimental techniques or conditions are required to observe and favour the appearance of the shell structure in unprocessed spectra. To begin with, the temperature of the source must be high enough to produce liquid clusters and avoid a production dominated by the geometric arrangement of the atoms (see Sect. 7.1). However, it should be stressed that this condition is not necessarily required in the very small size range or for elements in which the pseudopotential and crystalline field (see below) are weak, e.g., the alkali elements. (The discrete ionic structure acts as a perturbation which only moderately removes the degeneracy of levels separated by a wide gap.)

A first technique consists in favoring unimolecular evaporation, i.e., atom-by-atom evaporation, of the surface atoms of clusters, sometimes using in-flight laser heating [absorption of several photons with energy  $h\nu$  less than the one-photon ionisation threshold ( $h\nu < \text{IP}$ ) to avoid ionisation of the cluster], before their mass separation by ionisation at the input of a time-of-flight spectrometer (see Appendix H). Since the dynamics of sequential (atom-by-atom) evaporation is highly sensitive to the dissociation energy, which is closely correlated with the electronic shell structure, the magic numbers  $N_e$  survive relatively much longer than the numbers  $N_e + 1$ , which have a low mean binding energy.

The second technique is known as near-threshold photoionisation. Photoionisation is the ionisation process, i.e., removal of an electron from the neutral cluster, induced by absorption of a photon. Near-threshold photoionisation consists in the soft ionisation of the clusters, where the term ‘soft’ indicates that absorption of at most one photon causes ionisation. The photon has energy  $h\nu$  only slightly higher than the ionisation potential of the clusters. (Note that, since the ionisation potential depends on the size  $N$ , the laser frequency must be adjusted to the appropriate range of sizes.) Since the ionisation cross-section, i.e., the probability of ionisation, is very sensitive to the difference  $h\nu - \text{IP}(N)$  (increasing as this difference increases), it is clear that the mass spectrum obtained will reflect the shell structure, and a sudden increase in the signal will indicate occupation of a new electron level (magic number plus one).

Figure 7.15 clearly shows how temperature affects observation of atomic and electronic shell structures (see also Fig. 7.10). When the source is at room temperature, the photoionisation spectrum of aluminium clusters  $\text{Al}_N$  ( $N_e = 3N$ , since aluminium is a trivalent metal) reflects the production of clusters of mean octahedral shape (close-packed face-centered cubic structure). An oscillation corresponds to the transition from the octahedron of index  $k$  to the octahedron of index  $k + 1$  when 4 of the 8 faces of the octahedron have been covered. The oscillation is steady on a  $N_e^{1/3}$  scale, i.e., on a ‘mean’ radius  $R_{N_e}$  scale (see Appendix I). The oscillation period, of the order of  $\Delta N_e^{1/3} \approx 0.3$ , is specific to this geometry. The mass spectrum thus reflects the development of an atomic shell structure ( $T_s = 300 \text{ K}$ ). Heating the source to 600 K, a steady oscillation of period  $\Delta N_e^{1/3} \approx 0.6$ , roughly double the value, is observed, in



**Fig. 7.15.** Mass spectra of aluminium clusters  $Al_N$  obtained at two operating temperatures  $T_s$  of the pulsed laser source. Clusters are produced by a pulsed laser focused on an aluminium rod, which is vaporised locally [19]. The upper spectra are unprocessed, whilst the lower spectra are obtained by subtracting from the corresponding unprocessed spectrum its highly smoothed envelope and applying a further slight smoothing process. This gives a clearer view of the oscillations. *Left:*  $T_s = 300$  K. Steady oscillations with period  $\Delta N_e^{1/3} \approx 0.3$  reflect the development of an atomic shell structure, the mean cluster geometry being the one associated with octahedral symmetry. *Right:*  $T_s = 600$  K. Heating the source has caused (total or partial) melting of the clusters and the steady oscillations with period  $\Delta N_e^{1/3} \approx 0.6$  now reveal a stability dictated by the electronic shell structure

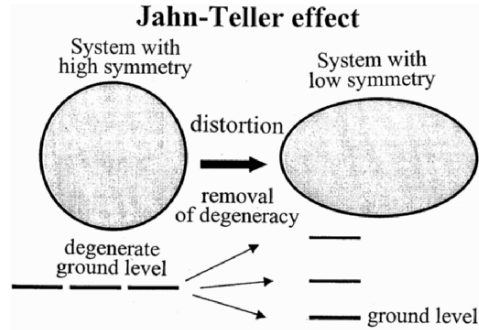
perfect agreement with the predictions of the jellium model. This modulation signals the production of liquid clusters whose stability is now governed by an electronic shell structure ( $T_s = 600$  K).

Due to the size dependence of the melting temperature of clusters as given by (7.43), viz.,

$$T = T_m = T_0 \left( 1 - \frac{a}{R} \right),$$

it has been possible to observe these two types of shell structure simultaneously in the mass spectrum of sodium clusters. In this case, the oscillations in the spectrum in the large size range are associated with cubo-octahedral or icosahedral symmetries, these having the same sequence of geometric magic sizes (the numbers  $N$  leading to perfect polyhedra).

Other information can be extracted from the mass spectrum in Fig. 7.14. Fluctuations are observed in the intensities of successive peaks, with period



**Fig. 7.16.** Illustration of the Jahn–Teller effect, a widespread phenomenon in different areas of physics. When a system has a high level of symmetry, e.g., spherical symmetry, as shown here, where the system is invariant under rotation, the quantum levels are almost always degenerate. A distortion (one speaks of symmetry breaking in physics) thus tends to increase the stability of the system

$\Delta N = 2$ , known as odd–even alternations, and to a lesser extent  $\Delta N = 4$ . These alternations also affect most of the physical properties of metallic clusters  $M_N$ . For example, in the case of monovalent elements, even neutral clusters ( $N = N_e$  even) are more abundant and hence more stable, and their ionisation potentials are higher than those of the adjacent odd sizes (see Fig. 7.14). As for the main effects due to electronic shells, the correlation between these fluctuations and the electronic structure of the nano-object is confirmed by a straightforward comparison with the properties measured for the ionic clusters  $M_N^+$  and  $M_N^-$ . Indeed, since the magic sizes correspond to specific numbers of electrons, the magic numbers  $N$  (number of atoms) are shifted by +1 or –1 in the mass spectra of cations ( $N_e = N - 1$ ) and anions ( $N_e = N + 1$ ), respectively, relative to the magic numbers  $N$  of the neutral clusters.

These fluctuations are associated with the loss of spherical symmetry induced by the so-called Jahn–Teller effect (see Fig. 7.16). This symmetry breaking is explained in an analogous way to the ellipsoidal deformations of non-magic atomic nuclei [where one must also include the Bardeen–Cooper–Schrieffer (BCS) nucleon pairing mechanism].

Let us give a brief description of the Jahn–Teller effect in the present case. Consider a non-magic size  $N_e$  whose electronic Fermi level  $E_{n,l}$ , obtained within the framework of a model which imposes spherical symmetry, such as the jellium model described earlier, is thus filled by a number of electrons  $m$  less than the degeneracy  $g = 2(2l + 1)$  of the one-electron level  $E_{n,l}$ . There are several ways of distributing these  $m$  electrons among the  $g$  individual states. Several states,  $p$  in number, therefore correspond to the ground level  $E_0$  of the system of  $N_e$  electrons. In fact,  $p$  is the number of different subsets containing  $m$  (distinct) elements that can be formed from a set containing  $g$  elements, i.e.,



$$p = \frac{g!}{m!(g-m)!}.$$

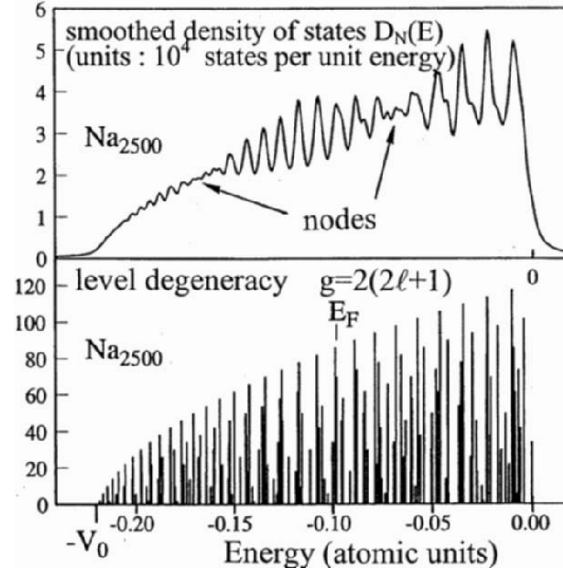
Distorting the cluster will have the effect of splitting the  $p$  times degenerate ground level  $E_0$  into several sublevels, some with energy less than  $E_0$  and others with energy greater than  $E_0$ . The ground state energy of the cluster is thus lowered when the symmetry is broken in this way (see Fig. 7.16), and minimal for a specific distortion. Using a model where there is no restriction of symmetry, such as a homogeneous jellium free to deform at constant volume (hence not necessarily spherical), the equilibrium shape of the cluster and its electronic structure can be determined. Considering an ellipsoidal distortion with axis of revolution  $Oz$ , the quantum number  $l$  specifying the one-electron levels  $E_{n,l}$  (leading to orbital degeneracy  $2l + 1$  for spherical symmetry) is replaced by the quantum number  $\Lambda = 0, \pm 1, \pm 2, \dots$ , associated with the projection of the angular momentum on the axis of revolution. (The invariance under rotations about the  $Oz$  axis is conserved despite the distortion.) This leads to orbital degeneracies of 2 for  $\Lambda \neq 0$  or 1 for  $\Lambda = 0$ , the energy depending only on  $|\Lambda|$ . The degeneracies of the one-electron levels are thus 4 or 2 when the spin degeneracy is included. This explains the observed odd–even alternations. The minimisation of the total energy with respect to the deformation parameters confirms the physical origin of the intensity fluctuations. The calculation also shows that, from one size to the next, the shape minimising the total energy can change radically, going from a cigar shape to a saucer shape, or vice versa, with the spherical shape being the equilibrium shape for the magic sizes ( $p = 1$ ).

### 7.4.2 Electronic Supershells

In this section, we shall first describe, then interpret, the electronic supershell structure. This structure is only observed in very large clusters  $N > 1000$  and corresponds to a periodic modulation in the intensity of the shell effects, i.e., depending on the size range, the highly stable magic sizes appear with different degrees of clarity in the mass spectra.

Is the shell structure observable in clusters of large size  $N$ ? A great many mechanisms or parameters which have little influence when  $N$  is small seem to limit observation of the shell structure to small cluster sizes. One may cite heat effects, shape isomerism, or the granular structure of the ionic background (these effects all lead to an effective broadening of the  $E_{n,l}$  levels by removing degeneracy), and more intrinsically, the decrease of the gaps  $\Delta E_{n,l}$  (compared with the energy  $k_B T$ ). A glance at the right-hand spectrum of Fig. 7.15 belies this presupposition.

In fact, when the spectra of the  $E_{n,l}$  levels are examined more closely, it turns out that they tend to bunch together to varying degrees depending on the energy range, suggesting that these bunches of levels will play the same role as the individual  $E_{n,l}$  levels in the small size range (see Fig. 7.17).



**Fig. 7.17.**  $E_{n,l}$  energy levels in a square well of depth  $V_0 = 0.22$  a.u.  $\sim 6$  eV and radius  $R = 53.34$  bohr, parameters corresponding to the cluster  $\text{Na}_{2500}$  [see Fig. 7.11 (square well) and Fig. 7.13 (value of  $V_0$ )]. The heights of the bars in the *lower graph* indicate the degeneracies of the individual electron levels, i.e.,  $g = 2(2l + 1)$ , and  $E_F$  is the Fermi level (the last level containing electrons). The energy spectrum is very dense for large values of the cluster size  $N$  (compare with Fig. 7.11). One can see, to differing degrees depending on the energy range, that the levels sometimes tend to bunch together. To bring out the non-uniformity of the distribution of quantum levels (more precisely, the quantum states) in the potential well, each infinitely thin bar with its peak at  $E_{n,l}$  is replaced by a curve of finite width, in this case, a Lorentzian curve of width  $\Delta = 0.01$  a.u. at half height, centered on  $E_{n,l}$ , whose integral is equal to  $g$ . (In mathematical terms, this is a convolution of the initial discrete density, represented by the bars, with a Lorentzian function.) The resulting spectrum  $D(E)$  is called the density of states and indicates the number of quantum states per unit energy. Of course, it depends on the degree of smoothing, i.e., the value of  $\Delta$ . This procedure brings out the main features in the non-uniformity of the distribution of states in the potential well. Note also that this level broadening procedure allows one to simulate physical effects such as the removal of level degeneracy when spherical symmetry is broken, or the inhomogeneous effects leading to a mean spectrum representing a set of clusters that are not strictly identical, and so on

When the spectrum is dense, this pattern is difficult to make out, but clear modulations are nevertheless visible in the smoothed density of states  $D_N(E)$  (see the caption to Fig. 7.17). The index  $N$  here reminds us that the density of states depends on  $N$ . A remarkable feature of this modulation is its quasi-harmonic (i.e., periodic) nature as a function of  $(E + V_0)^{1/2}$ , where  $E + V_0$  is the energy measured from the bottom of the potential well, which is just

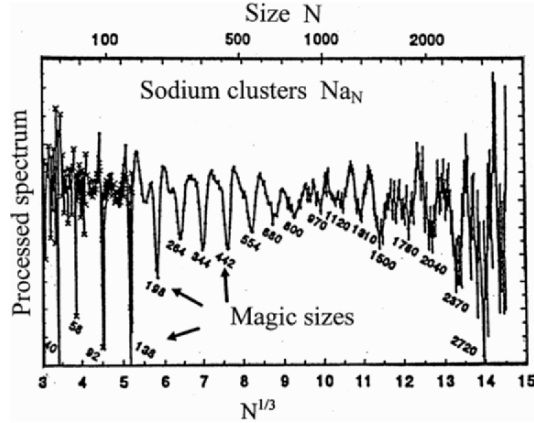
the kinetic energy  $E_{\text{kin}}$ . The properties of the clusters are directly correlated with the electronic spectrum  $E_{n,l}$  in the vicinity of the Fermi level  $E_{\text{F}}(N)$ , and hence also with  $D_N(E_{\text{F}})$ . Note that  $E_{\text{F}}(N)$  is almost independent of the size (see Fig. 7.13), and indeed, the variations are tiny for large sizes  $N \gtrsim 10^2$ . Owing to the approximate scaling law  $E_{n,l} \propto 1/R_{N_e}^2$ , which squeezes the modulations together at the bottom of the well as the size  $N_e$  increases, this new type of shell structure, described in terms of bunches of levels or oscillations in the density of states, is expected to show up directly in experimental mass spectra, but also in the dependence of the physical properties on cluster size.

This theoretical prediction has been confirmed by experiment. Mass spectra with steadily oscillating intensity as a function of  $N_e^{1/3}$ , analogous to the right-hand spectrum of Fig. 7.15, have been obtained for most elements which have clearly identifiable magic sizes in the small size range. The observed period  $\Delta N_e^{1/3} \approx 0.6$  is often said to be universal, because it is the same for all metals.

Another remarkable feature of the experimental spectra, and indeed the smoothed densities of states  $D(E)$ , is the electronic supershell structure, which corresponds to a large scale periodic modulation in the amplitude of the shell effects, analogous to the beat pattern resulting from interference between two signals with slightly different frequencies. This beat pattern is characterised in particular by a shift of one half period ( $\Delta N_e^{1/3} \approx 0.3$ ) in the periodic sequence of the maxima (or minima) of the modulated spectrum as one crosses the zone where the shell effects are almost non-existent. This zone is called the node (see Figs. 7.17 and 7.18). On either side of the node, the consecutive magic sizes are separated by a distance of  $\Delta N_e^{1/3} \approx 0.6$  on the  $N_e^{1/3}$  scale, but between the two series, there is a further shift equal to  $\Delta N_e^{1/3} \approx 0.3$ . As we shall see later, this superstructure, described in terms of oscillations in the density of states  $D(E)$ , results from a kind of interference between two energy structures with very similar periods. All the features described here will be interpreted using the semiclassical theory of the density of states.

The first clear experimental confirmation of this supershell structure was obtained with sodium clusters. The sensitivity to the relative stability was enhanced by sequential unimolecular evaporation during free flight (see Fig. 7.18).

From the unprocessed spectrum, almost smooth to the eye above  $N = 500$  atoms for the reasons explained at the beginning of this section, the beat node was visualised by applying sophisticated signal processing techniques. This beat node, observed at around 1 000 electrons, is accompanied by a shift of one half period in the sequence of magic sizes. This phenomenon had been anticipated theoretically for some time, particularly in the context of the shell structure in atomic nuclei, hence the term ‘confirmation’ used above. Of course, the small number of nucleons ( $< 250$ ) makes direct experimental demonstration impossible. It should be stressed that this superstructure is not a property specific to a gas of delocalised conduction electrons in metallic



**Fig. 7.18.** Signal obtained after processing the mass spectrum of sodium clusters to account for temperature and size effects, revealing the electronic supershell structure [20]. Magic sizes correspond to minima of the spectrum. The amplitude of the shell effects goes through a minimum near  $N = 1000$ , a zone known as the node of the electronic supershell structure. A shift of  $\Delta N^{1/3} \approx 0.3$  in the periodic sequence of successive magic sizes is also observed at the beat node

clusters, but is in fact a property of the energy spectrum of a system of spatially confined fermions, when the effective confining potential  $V_{\text{eff}}(r)$  has the form illustrated in Figs. 7.11 and 7.13.

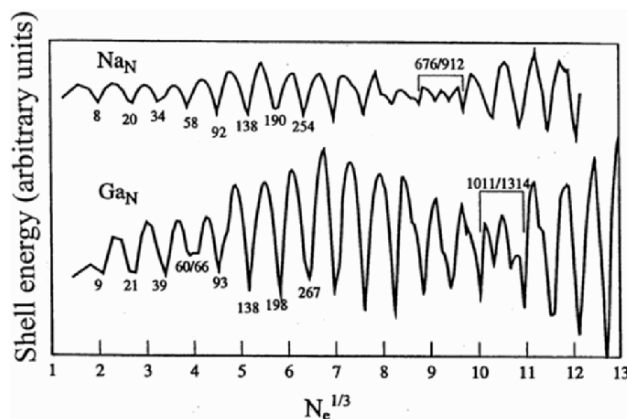
The predictions made using the jellium model for sodium (monovalent,  $N_e = N$ ,  $r_s = 3.93$  bohr) and gallium (trivalent,  $N_e = 3N$ ,  $r_s = 2.19$  bohr) are shown in Fig. 7.19. The quantity calculated is the shell energy  $E_{\text{shell}}(N_e)$ , the quantum contribution to the energy, defined by the relation

$$E_0(N_e) = E_{\text{ld}}(N_e) + E_{\text{shell}}(N_e), \quad (7.88)$$

where  $E_0(N_e)$  is the total energy of the cluster and  $E_{\text{ld}}(N_e)$  is the classical liquid-drop-type contribution (a third degree polynomial in  $N_e^{1/3}$ ). It should be noted that, for large clusters,  $E_{\text{shell}}(N_e)$  only makes a tiny contribution to  $E_0(N_e)$ . The shell effects in gallium, characterised by small  $r_s$  compared with the alkali elements, are much more significant than in sodium. This is a consequence of the approximate scaling law

$$E_{n,l} \propto 1/R_{N_e}^2,$$

obeyed by the individual levels as the size increases (larger gaps between successive levels). In fact, gallium is the only metal with a low enough melting temperature ( $T_m = 303$  K) for which the supershell structure has been directly visible in the mass spectra. Although the  $\Delta N_e^{1/3} \approx 0.6$  periodicity in the sequence of successive magic numbers appears to be a universal attribute of the effective potentials  $V_{\text{eff}}(r)$  with flat bottom and steep enough walls (see

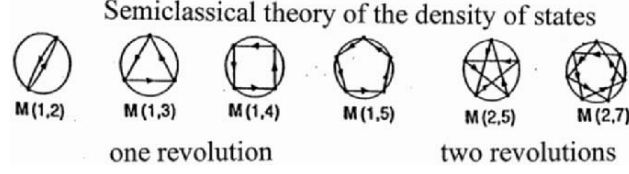


**Fig. 7.19.** Shell energy calculated using the jellium model for sodium clusters (monovalent,  $N = N_e$ ) and gallium clusters (trivalent,  $N_e = 3N$ ), exhibiting the electronic supershell structure. The shell energy, the quantum contribution to the energy, is obtained by subtracting the classical contribution of liquid-drop type, expressed in the form of a power series in  $N_e^{1/3}$ , from the total energy of the cluster. The smallest magic numbers are indicated (sizes given by the number of conduction electrons). The *horizontal dash* bounded by two large magic numbers indicates the region of the beat node in the electronic supershell structure, i.e., the zone in which the amplitude of shell effects is minimal

Figs. 7.11 and 7.13), the differences observed in the magic numbers (of course, in the present case, there is also the fact that the number of electrons is a multiple of 3 for gallium) and the positions of the beat node (depending on the metal, except for the magic numbers in the very small size range) can be attributed almost exclusively to the relative influence of the profile of  $V_{\text{eff}}(r)$  on the quantisation of the energy levels  $E_{n,l}$  (for fixed size  $N_e$ , this influence increases as  $r_s$  decreases).

### Semiclassical Interpretation of Electronic Shell and Supershell Structures

The main features of the electronic shell and supershell structures, and in particular the universal period  $\Delta N_e^{1/3} \approx 0.6$  and the beat pattern, can be interpreted within the more intuitive framework of the semiclassical theory of the density of states of an electron confined by a potential well  $V_{\text{eff}}(r)$ . Note that this theory is closely related to the semiclassical quantisation rules derived from an action integral of Bohr–Sommerfeld type. The semiclassical formalism is the only way of ‘explaining’ the shell structure for the large size range, e.g., explaining why consecutive magic sizes arise periodically on the  $N_e^{1/3}$  scale, with period  $\Delta N_e^{1/3} \approx 0.6$ , and of providing an interpretation of the electronic supershell structure. Moreover, it provides a useful intuitive



**Fig. 7.20.** Closed classical trajectories of an electron confined in a spherical potential well with flat bottom and vertical walls (see the left-hand diagram of Fig. 7.11). Only the shortest trajectories with one or two revolutions are depicted.  $M(q, n)$  indicates that the trajectory is characterised by  $q$  revolutions and  $n$  rebounds against the surface

guide and predictions can be formulated without always going through laborious quantum calculations. This feature, which is always appreciable when discussing quantum effects, will be illustrated when we discuss the influence of the profile of the potential  $V_{\text{eff}}(r)$  on the electronic supershell structure at the end of this section. We shall only present those results required to understand this discussion.

In this theory, it is shown that the smoothed density of states  $D(E)$  can be written as the sum of a slowly varying function of  $E$  (an expansion of liquid-drop type containing bulk, surface and curvature terms) and a fluctuating contribution expressed in terms of the action integrals of the closed classical orbits of the electron (of energy  $E$ ), viz.,

$$D(E) = D^{\text{smooth}}(E) + D^{\text{fluc}}(E), \quad (7.89)$$

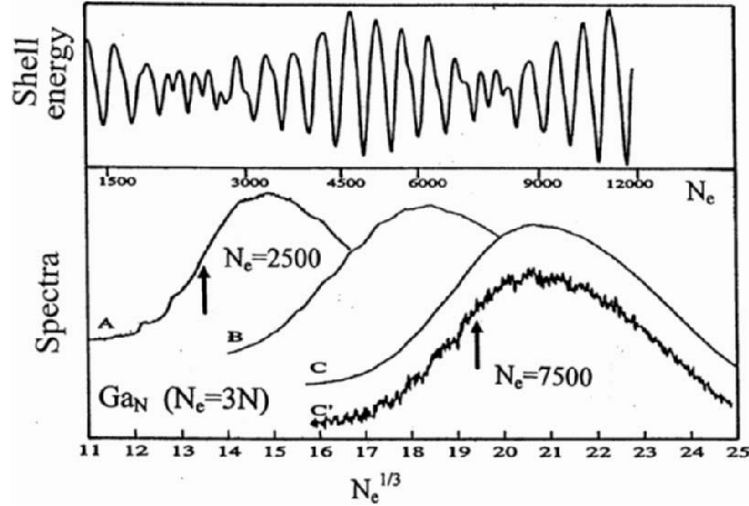
with

$$D^{\text{fluc}}(E) = \sum_{M(q,n)} A_M \cos \left[ \frac{1}{\hbar} \oint_M \mathbf{p}(\mathbf{r}) d\mathbf{r} + \Psi_M \right], \quad (7.90)$$

where  $M(q, n)$  specifies a closed classical trajectory with  $n$  radial oscillations (or  $n$  rebounds against the outer surface of the effective potential) and  $q$  revolutions. The quantities  $\mathbf{p}(\mathbf{r}) = m\mathbf{v}(\mathbf{r})$  and  $\mathbf{v}(\mathbf{r})$  are the momentum and velocity of the electron, respectively, which depend on its position  $\mathbf{r}$  as it moves around its orbit. Finally,  $A_M$  and  $\Psi_M$  are the amplitude and phase factors of the contribution from the closed orbit labelled by  $M$ .

Figure 7.20 illustrates the shortest trajectories for  $q = 1$  (regular polygons) and  $q = 2$  (stars with 5, 7, etc., points) with a different topology, i.e., not just polygonal trajectories covered several times, of the form  $M(q = 2, n' = 2n)$ , in the case of a spherical well with flat bottom and vertical walls, corresponding to the left-hand diagram of Fig. 7.11.

To simplify, we begin by considering the case of a strictly square-shaped potential (in 3D) with radius  $R = r_s N_e^{1/3}$  and depth  $V_0$ , like the one illustrated in Fig. 7.11. The action integral is easily calculated and can be expressed as



**Fig. 7.21.** *Lower:* Mass spectra of gallium clusters (trivalent,  $N_e = 3N$ ) obtained by near-threshold photoionisation using a laser vaporisation source. The spectrum C' was obtained by subtracting 80% of the highly smoothed envelope from the spectrum C. *Upper:* Shell energy calculated using the jellium model, taking into account the effects of the electron-ion interaction pseudopotentials and assuming a non-step-like ionic surface distribution (width 3 bohr). The first two nodes (where the amplitude of shell effects is minimal) in the electronic supershell structure are observed near  $N_e = 2500$  and  $7500$  electrons [21]

the product of the momentum (independent of the position on the orbit) and the length of the trajectory:

$$\oint \mathbf{p}(\mathbf{r}) d\mathbf{r} = (2mE_{\text{kin}})^{1/2} L_M = (2mE_{\text{kin}})^{1/2} \alpha_M r_s N_e^{1/3}, \quad (7.91)$$

where  $E_{\text{kin}} = E + V_0$  is the kinetic energy of the electron, so that  $p = (2mE_{\text{kin}})^{1/2}$ , and  $L_M = \alpha_M R$  the length of the trajectory, with  $\alpha_M$  a number specific to the trajectory labelled by  $M$ , e.g.,  $\alpha_s = 4 \times 2^{1/2}$  for the square,  $\alpha_t = 3 \times 3^{1/2}$  for the triangle, and so on. In the case of real effective potentials like the ones in Fig. 7.11 (Woods-Saxon-type potential) and Fig. 7.13 (self-consistent potential calculated using the jellium model), the trajectories  $M(q, n)$  are no longer made up of straight line segments traversed at constant speed, but tend to be distinctly curved at the rebounds from the surface.

However, the calculation shows that the action integral closely follows a scaling law of type

$$\oint \mathbf{p}(\mathbf{r}) d\mathbf{r} = (2mE_{\text{kin}})^{1/2} \alpha_M r_s N_e^{1/3} + \Omega_M, \quad (7.92)$$

where  $\Omega_M$  is a correction to the simple case considered above, accounting for fluctuations in the bottom of the well around the mean value  $-V_0$  (leading to

slight fluctuations in the momentum about its mean value) and also effects due to the edge of the potential [modification of the length of the trajectory and changes in the electron speed  $v(\mathbf{r})$  in the rebound zones]. The remarkable thing is that  $\Omega_M$  is practically independent of the cluster size. This happens because the profile of  $V_{\text{eff}}(r)$  is itself independent of the size, and because the trajectory is essentially contained in the core of the cluster [ $V_{\text{eff}}(r) \approx -V_0$ ]. In the following, we set

$$\Phi_M = \Psi_M + \frac{1}{\hbar} \Omega_M . \quad (7.93)$$

Given the above equations, the semiclassical expansion of the density of states can be considered as the Fourier decomposition of  $D(E_{\text{kin}}^{1/2})$ , with each of the harmonic components  $D^M(E_{\text{kin}}^{1/2})$  associated with one closed orbit of the electron.

It is now possible to interpret the regular modulations observed in the mass spectra. From the sensitivity of the physical observables to  $D_N(E_F)$  (the index  $N$  reminds us that the density of states depends on  $N$ ) and the weak dependence of  $E_F$  on  $N$  ( $E_{\text{cF}}$  fluctuates around the value corresponding to an electron gas with charge density  $-n_0$ ), i.e.,

$$E_{\text{cF}} = \frac{\hbar^2}{2m} \left( \frac{3\pi^2 n_0}{q} \right)^{2/3} , \quad (7.94)$$

it follows that each of the harmonic components  $D_N^M(E_{\text{kin}}^{1/2})$  induces a modulation in the properties of the clusters which is periodic in  $N_e^{1/3}$ . It can also be shown that the density  $D_N^{\text{fluc}}(E)$  is dominated by the contributions from the shortest trajectories, mainly the square and triangular orbits, which describe overall oscillating structures in the smoothed density of states  $D(E)$  (see Appendix J).

Since the amplitudes  $A_M$  of the triangular and square trajectories are of the same order of magnitude, the electronic supershell phenomenon is a consequence of the beat pattern resulting from interference between the contributions from the triangular and square orbits, which have similar lengths (with a relative difference of about 10%). This leads to the result

$$D_{N_e}^{\text{fluc}}(E_F) \approx 2A \cos \left[ a \left( \frac{\alpha_s - \alpha_t}{2} \right) N_e^{1/3} + \frac{\Phi_s - \Phi_t}{2} \right] \times \cos \left[ a \left( \frac{\alpha_s + \alpha_t}{2} \right) N_e^{1/3} + \frac{\Phi_s + \Phi_t}{2} \right] , \quad (7.95)$$

where  $a = (9\pi/4)^{1/3}$ .

This expression leads to a simple explanation for all the experimental observations. The second oscillating factor (the ‘fast’ term) is responsible for shell effects periodically modulating the mass spectra, with a period of the



order of  $\Delta N_e^{1/3} \approx 0.6$ . The first factor, governing the amplitude of the shell effects, which has longer period (the ‘slow’ term), is responsible for the supershell structure. Nodes correspond to size ranges for which the phase of the first oscillating factor is close to  $\pi/2$ ,  $3\pi/2$ , etc. The shift  $\Delta N_e^{1/3} = 0.3$  in the regular sequence of successive magic numbers, observed at the nodes of the supershell structure, results from the fact that this first factor changes sign at each node: the logical continuation of the maxima appear as minima and vice versa each time a node is crossed.

The shell and supershell structures are clearly visible in the spectra of gallium clusters obtained by near-threshold photoionisation (see Fig. 7.21). However, the first two nodes of the supershell structure, observed near  $N_e = 2\,500$  and  $N_e = 7\,500$  electrons, disagree totally with the predictions of the jellium model, which gives  $N_e = 1\,150$  and  $N_e = 4\,500$  electrons. The underestimate of the position of the nodes in the jellium model is in fact systematic, as suggested by comparing the data for sodium clusters shown in Figs. 7.18 and 7.19.

It is easy to explain this disagreement. In the small size range, this simple model correctly predicts the magic sizes, because the gaps between successive levels or groups of levels are relatively large. But for large cluster sizes, the electronic spectrum is extremely dense (see Fig. 7.17), and the level bunching pattern, which can only be analysed via the smoothed density of states  $D(E)$ , then becomes highly sensitive to the profile of the effective potential  $V_{\text{eff}}(r)$  which governs the conditions for quantisation of energy levels. Hence, in order to interpret the shell and supershell structures in the large size range, further physical ingredients must be introduced into the jellium model, e.g., the effects of the electron–ion interaction pseudopotential (see Chap. 18) and the effects of a possible softness in the ionic surface density.

The main effect of these new ingredients is to soften the profile of  $V_{\text{eff}}(r)$  and hence to modify the position of the beat nodes. This happens because the closed trajectories are rounded off where they rebound from the surface, thereby lengthening them and modifying the phase shift  $\Phi_M$  in (7.93) and (7.95).

### 7.4.3 Optical Properties. Collective Excitations

A monochromatic electromagnetic wave is a useful tool for probing the electronic structure of clusters, just as it is for atoms or molecules. In the case of nanoscale clusters, optical properties constitute a field of investigation in its own right. Optical properties of clusters are an extremely rich subject of study due to their specific characteristics and their potential applications in optoelectronics and nanooptics. In this section, we shall consider only metallic clusters, which exhibit the most pronounced quantum finite size effects, and we shall only discuss their absorption and polarisation properties. Before illustrating these properties and interpreting the observed quantum effects, we shall outline the optical properties predicted by the classical theory as a way

of introducing the main ideas. Throughout the section, the complex dielectric functions  $\varepsilon(\omega) = \varepsilon_1(\omega) + i\varepsilon_2(\omega)$  are the dimensionless quantities defined relative to the vacuum permittivity  $\varepsilon_0$  (see Appendix K).

As the number of atoms  $N$  in the cluster increases, the transition from an atomic- or molecular-type spectrum towards the spectrum of a small piece of matter, described classically, happens very quickly. A small cluster nevertheless exhibits a very different optical response to a macroscopic solid due to dielectric confinement, which gives rise to collective electronic excitations in the gas of delocalised conduction electrons. Such collective excitations are also called Mie resonances or plasmons. The term ‘dielectric confinement’ refers to classical finite size effects in the context of optical properties. In addition to these classical confinement effects, we shall find that quantum finite size effects must be included, due to quantisation of the electronic levels. These giant resonances, analogous to collective excitations in atomic nuclei, focus practically all the oscillator strength of the interaction with the electromagnetic field, generally in the visible to near UV region of the spectrum (energies of the order of a few eV), which is why the term ‘giant resonance’ is used.

The gradual appearance of these collective modes and their relative importance in the scattering and absorption properties are mainly governed by the ratio of the cluster radius  $R$  to the wavelength  $\lambda$  of the incident light. The classical Mie theory based on solution of Maxwell’s equations and a multipole expansion of the incident and scattered fields accounts perfectly for this size dependence. For nanoscale clusters (radius  $R \lesssim 20$  nm), the response is dominated by the dipole absorption term, which corresponds to the electric dipole approximation, neglecting propagation effects. In this approximation, the spatial dependence of the incident field on the scale of the nano-object can be ignored. In fact,  $\lambda \ll R$ , and all electrons in the nano-object thus feel the same oscillating field

$$\mathbf{E}(t) = \text{Re}(\mathbf{E}_0 e^{-i\omega t})$$

in complex notation, at any instant of time. The determination of this term thus reduces to a simple electrostatic calculation, whence the rather misleading description ‘quasi-static regime’ to qualify the electric dipole approximation.

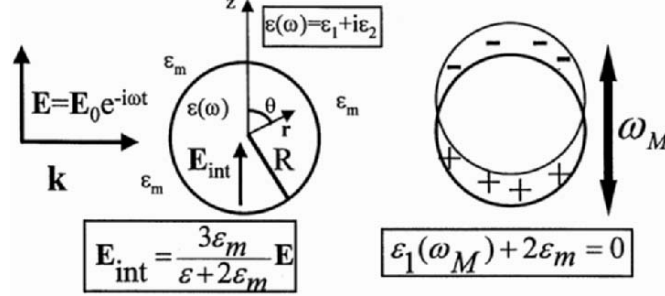
Figure 7.22 defines the various parameters of the problem. In the presence of an applied electric field

$$\mathbf{E}(t) = \mathbf{E}_0 e^{-i\omega t},$$

assumed to be polarised along the  $Oz$  axis, the time dependences of the various potentials and fields (in complex notation) arising in the problem are all of the form  $e^{-i\omega t}$ , e.g.,

$$V(\mathbf{r}, t) \equiv V(\mathbf{r}) e^{-i\omega t}.$$

$V(\mathbf{r})$  is obtained by solving the Poisson equation



**Fig. 7.22.** Surface plasmon resonance (Mie resonance) of a metallic cluster [complex dielectric function  $\varepsilon(\omega)$ ] embedded in a transparent matrix with real dielectric function  $\varepsilon_m(\omega)$ , i.e., the matrix is assumed to be non-absorbing. For many common matrices, such as alumina, silica, aqueous bath, etc.,  $\varepsilon_m(\omega)$  depends very weakly on  $\omega$  in the visible/near UV region of the electromagnetic spectrum (in which the surface plasmon excitation is observed). This is why the angular frequency dependence of  $\varepsilon_m(\omega)$  is left out for simplicity. The angular frequency  $\omega_M$  of the Mie resonance, which corresponds to the collective oscillation of the conduction electrons relative to the positively charged ionic background is an approximate solution to the equation  $\varepsilon_1(\omega_M) + 2\varepsilon_m = 0$ , where  $\varepsilon_1(\omega)$  is the real part of the complex dielectric function of the metal.  $\mathbf{E}$  is the applied oscillating field and  $\mathbf{E}^{\text{int}}$  is the (spatially homogeneous) internal oscillating field within the metallic cluster

$$\Delta V(\mathbf{r}) = 0$$

in spherical coordinates, with the following boundary conditions:

- $V(\mathbf{r}) = -E_0 r \cos(\theta)$  as  $r \rightarrow \infty$ ,
- the electrostatic potential  $V(\mathbf{r})$  is continuous at the metal/matrix interface,
- the normal component of the displacement vector  $\mathbf{D}(\mathbf{r}) = \varepsilon_0 \varepsilon(\omega) \mathbf{E}_{\text{tot}}(\mathbf{r})$  is continuous at the interface, where  $\mathbf{E}_{\text{tot}}(\mathbf{r})$  is the total macroscopic electric field, the sum of the applied field and polarisation fields.

It then follows that

$$V(\mathbf{r}) = \left( \frac{\varepsilon - \varepsilon_m}{\varepsilon + 2\varepsilon_m} \frac{R^3}{r^3} - 1 \right) E_0 r \cos(\theta) = -\mathbf{E}_0 \cdot \mathbf{r} + \frac{1}{4\pi\varepsilon_0} \frac{\mathbf{p} \cdot \mathbf{r}}{r^3}, \quad (7.96)$$

for  $r > R$ , and

$$V(\mathbf{r}) = \frac{-3\varepsilon_m}{\varepsilon + 2\varepsilon_m} E_0 r \cos(\theta) = -\mathbf{E}^{\text{int}} \cdot \mathbf{r}, \quad (7.97)$$

for  $r < R$ .

The electric field  $\mathbf{E}^{\text{int}}$  is thus uniform inside the cluster and the external field is the sum of the incident field  $\mathbf{E}(t)$  and the field created by a fictitious point dipole  $\mathbf{p}(\omega, t)$  placed at the center of the cluster such that

$$\mathbf{p}(\omega, t) = 4\pi\varepsilon_0 \frac{\varepsilon - \varepsilon_m}{\varepsilon + 2\varepsilon_m} R^3 \mathbf{E}(t) = \alpha(\omega) \mathbf{E}(t) . \quad (7.98)$$

The dynamic polarisability  $\alpha(\omega)$  is the main quantity characterising the optical properties of the cluster ( $R/\lambda \ll 1$ ). The static polarisability  $\alpha(0) = 4\pi\varepsilon_0 R^3$  ( $|\varepsilon| \gg \varepsilon_m$  at very low frequencies) is therefore a measure of the volume of the cluster.

The mean power  $P_{\text{mean}}$  dissipated inside the cluster can be expressed in terms of the imaginary part of the dynamic polarisability (component in phase quadrature with the incident field):

$$P_{\text{mean}} = \left\langle \int_{\text{vol}} \text{Re}[\mathbf{E}^{\text{int}}(t)] \cdot \frac{d}{dt} \text{Re}[\mathbf{P}^{\text{int}}(t)] d^3r \right\rangle = \frac{E_0^2 \omega}{2} \varepsilon_m \text{Im}[\alpha(\omega)] . \quad (7.99)$$

The absorption cross-section  $\sigma(\omega)$ , with units of length squared, is obtained by dividing  $P_{\text{mean}}$  by the energy flux incident in the matrix, i.e.,

$$I_0 = \frac{1}{2} [E_0^2 c (\varepsilon_m)^{1/2} \varepsilon_0] . \quad (7.100)$$

From (7.99) and (7.100), we obtain

$$\sigma(\omega) = \frac{P_{\text{mean}}}{I_0} = \frac{\omega \varepsilon_m^{1/2}}{c \varepsilon_0} \text{Im}[\alpha(\omega)] = \frac{9\omega \varepsilon_m^{3/2}}{c} \frac{4\pi R^3}{3} \frac{\varepsilon_2}{(\varepsilon_1 + 2\varepsilon_m)^2 + (\varepsilon_2)^2} \propto R^3 . \quad (7.101)$$

Note that, apart from the multiplicative volume factor, the classical theory predicts no finite size effect in the electric dipole approximation. In general,  $\varepsilon_2(\omega)$  (the imaginary part of the dielectric function of the metal) is relatively small and varies very little in the relevant spectral range, and a very high level of absorption is observed in the region minimising the denominator of the last expression. This leads to the approximate resonance condition

$$\varepsilon_1(\omega_M) + 2\varepsilon_m(\omega_M) = 0 ,$$

which defines the collective excitation frequency  $\omega_M$ . This resonance, known as the surface plasmon resonance, or Mie resonance, corresponds classically, and indeed in quantum physics, to the collective oscillation of the conduction electrons relative to the positively charged ionic background (see Fig. 7.22). Such a collective excitation can be compared with the giant resonance observed in atomic nuclei, associated with a relative oscillation between the proton and neutron fluids.

Let us now make a more accurate estimate of the resonance frequency for simple metals like the ones cited in previous sections (alkali metals, noble metals). As a first approximation,  $\varepsilon(\omega)$  can be decomposed into a contribution  $\varepsilon^s(\omega)$  of Drude–Sommerfeld type, describing the optical properties of the gas of conduction electrons, and a contribution

$$\varepsilon^d(\omega) = \varepsilon_1^d(\omega) + i\varepsilon_2^d(\omega) ,$$

characterising the optical properties of electrons in the ionic cores. The so-called ionic contribution, which corresponds to collective displacements of the ions relative to one another, is only relevant in the far infrared, i.e., vibration periods of the order of the picosecond, and can be neglected here. We then obtain

$$\varepsilon(\omega) = \varepsilon^s(\omega) + \varepsilon^d(\omega) - 1 = 1 + \chi^s(\omega) + \chi^d(\omega) , \quad (7.102)$$

with

$$\varepsilon^s(\omega) = 1 - \frac{\omega_p^2}{\omega(\omega + i\Gamma)} , \quad (7.103)$$

where  $\omega_p = [n_0q/\varepsilon_0m]^{1/2}$  is the angular frequency of the bulk plasma of the metal and  $\Gamma$  a damping constant characterising all collision processes affecting the conduction electrons (vibrations, defects, impurities, etc.). For such metals, the resonance angular frequency  $\omega_M$  is an approximate solution of the implicit equation

$$\omega_M = \frac{\omega_p}{[\varepsilon_1^d(\omega_M) + 2\varepsilon_m]^{1/2}} , \quad (7.104)$$

or approximately,

$$\omega_M = \frac{\omega_p}{[1 + 2\varepsilon_m]^{1/2}} \quad (7.105)$$

for the alkali metals, since  $\varepsilon^d(\omega) \approx 1$ . In this simple case, it is easy to show that the absorption spectrum given by (7.101) is approximately a Lorentzian curve centered on  $\omega_M$  with width  $\Gamma$  at half-maximum. Hence,

$$\sigma(\omega) \propto \frac{1}{[(\omega - \omega_M)^2 + (\Gamma/2)^2]} .$$

A direct determination of the absorption cross-section of clusters in the gas phase, e.g., by measuring the light transmission, would seem to be impossible at the present time because the cluster density in beams produced by standard sources is too low. This problem has led to the development of photoevaporation spectroscopy, a technique which we shall now outline.

A first variant known as photodepletion consists in measuring the intensity variation in a beam of clusters  $M_N$ , tightly collimated using diaphragms, after absorption of a photon. The energy  $h\nu$  of the photon is first transferred to the electrons and then rapidly converted into heat (vibrations of the ions). This causes one or more atoms to evaporate and the recoil effect then ejects fragments  $M_{N-p}$  from the collimated beam. This approach is based on the assumption of perfect proportionality between the absorption cross-section and the drop in intensity of the beam.

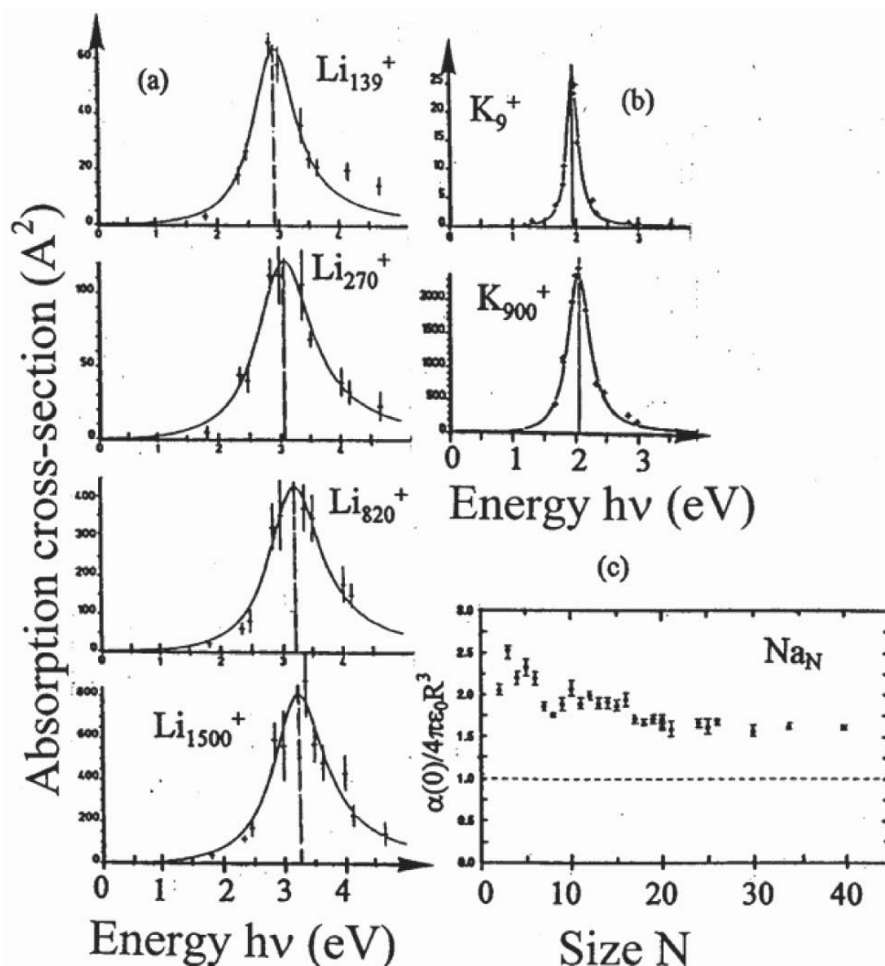
A second variant consists in analysing the distribution of fragments produced by sequential evaporation following the absorption of several photons, using a dynamic evaporation model (usually, statistical).

It should be emphasised that these two techniques are both indirect, in the sense that they are based on the assumed validity of some dynamic evaporation model. This is why most experimental work in the solid phase (as opposed to methods involving clusters dispersed in solution) uses the codeposition technique. Here, clusters are embedded in low concentrations in a transparent dielectric matrix, vaporised and deposited simultaneously with the clusters on a transparent substrate, in the form of a composite thin film. The properties of the film are then determined by conventional optical techniques, such as transmission or ellipsometry. In this section, we shall not discuss the methods developed in the colloid community or by chemists, where the nano-objects are very often grown in situ. The thin films produced in this way can be processed, e.g., by annealing, and characterised later using complementary techniques such as microscopy, Rutherford back-scattering of ions, X-ray diffraction and scattering, etc. This provides a full description of the relevant nano-objects and a reliable means of correlating optical and morphological properties. However, these approaches also suffer from a number of limitations:

- The geometry and electronic structure of the free cluster may not be preserved when it is embedded in a matrix (the so-called memory effect).
- The measured optical properties depend on the environment via  $\varepsilon_m(\omega)$  (see previous formulas), and we shall see later that they are sensitive to features of the metal/matrix interface such as defects, chemical bonds, and local porosity. It is therefore essential to take these factors into account, from the fundamental standpoint, if we wish to extract the intrinsic properties of the clusters from the optical measurements.

Regarding the static polarisability  $\alpha(0)$  (or the permanent dipole moment of the nano-object), which provides information concerning the properties of the ground electronic state, such as the ‘volume’ of the gas of conduction electrons in the case of a homogeneous cluster, charge transfer within more complex nano-objects possessing a permanent dipole moment, e.g., inhomogeneous mixed clusters or clusters formed by stacking several smaller entities, the free phase technique involves measuring the deflection of clusters as they cross an inhomogeneous electric field with a steep gradient. This is analogous to the Stern–Gerlach experiment used to measure the magnetic moments of atoms.

Figure 7.23 shows results obtained by various groups with alkali elements in the gas phase ( $\varepsilon_m = 1$ ), using photoevaporation spectroscopy (absorption cross-sections of charged clusters  $K_N^+$  and  $Li_N^+$ ) and deflection in an inhomogeneous electric field with large gradient (static polarisability of  $Na_N$  clusters). These graphs provide conclusive, and unexpected, evidence that the optical properties are qualitatively similar to the predictions of the classical theory,



**Fig. 7.23.** (a) and (b). Absorption spectra of potassium clusters  $\text{K}_N^+$  and lithium clusters  $\text{Li}_N^+$  obtained by photoevaporation spectroscopy, i.e., analysing the distribution of fragments. The surface plasmon resonance is clearly visible. The experimental results are fitted to a Lorentzian curve [22, 23]. (c) Static polarisabilities of sodium clusters  $\text{Na}_N$ , normalised with respect to the classical value  $\alpha_{\text{cl}}(0) = 4\pi\epsilon_0 R^3$ , determined by measuring the deflection of clusters as they pass through an inhomogeneous electric field with large gradient [24]

even for very small cluster sizes. It should be remembered that the classical theory is only applicable a priori to macroscopic objects, e.g., the dielectric function of a metal directly reflects its band structure.

However, quantitatively, large discrepancies are observed, and these discrepancies grow smaller as the size  $N$  increases. The value of the static polarisability, which is a measure of the volume of the cluster, is much greater

than the classically predicted value. A quantitative analysis of the mean size dependence leads to a law of type

$$\alpha(\omega = 0, R_{N_e}) = 4\pi\epsilon_0(R_{N_e} + \delta)^3, \quad (7.106)$$

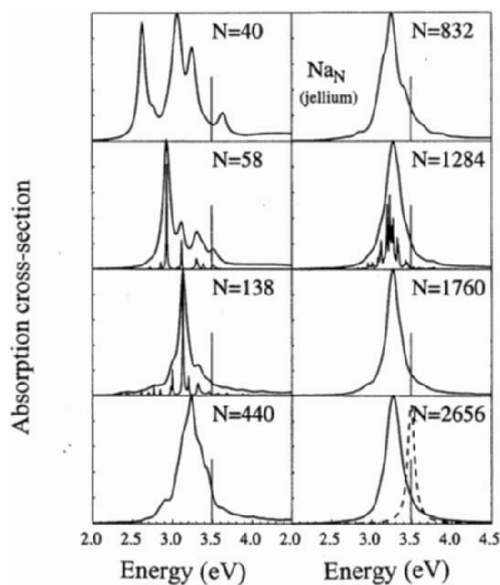
where  $\delta$  is a constant parameter. The observed absorption band, which corresponds to the collective dipole resonance, is centered on a much lower frequency than the value predicted by the classical theory. (We speak then of a redshift.) A Drude–Sommerfeld-type parametrisation of the dielectric functions of potassium and lithium ( $r_s \approx 4.86$  bohr and  $r_s \approx 3.25$  bohr, respectively) leads to the values  $\omega_M(\text{K}) = 2.54$  eV and  $\omega_M(\text{Li}) = 4.64$  eV. When the experimentally measured dielectric functions are used, the discrepancies are reduced, in particular for lithium [ $\omega_M(\text{K}) = 2.22$  eV and  $\omega_M(\text{Li}) = 3.55$  eV], showing that other physical effects, not accounted for in the simple Drude–Sommerfeld model, must be introduced to obtain a correct description of the optical properties of metallic clusters.

A simple interpretation of the significant redshift due to finite size quantum effects, and its mean size dependence, will be given below. But first, we must examine the quantum theoretical predictions.

Figure 7.24 shows results obtained with the time-dependent local density approximation (TD LDA) (discussed further in Chap. 18) for magic clusters of sodium in vacuum ( $\epsilon_m = 1$ ), using the jellium model. As in the classical theory, this approach is self-consistent insofar as the induced polarisation  $\mathbf{P}(\mathbf{r})$  at a point  $\mathbf{r}$  depends not only on the incident field  $\mathbf{E}(t)$ , but also on the fields created by the polarisation of the various media, i.e., the metallic cluster and matrix. This formalism also accounts for all quantum finite size effects ignored in the classical approach. Clearly, some effects not included in the simple jellium model, such as effects due to the polarisation of the ionic cores, are still not represented. The reader interested in the theoretical basis of the TD LDA formalism is referred to the literature [12–15]. As the integral  $\int \sigma(\omega)d\omega$  of the absorption spectrum is proportional to the size  $N$ , an arbitrary scale has been chosen. Moreover, only the spectral range in which the oscillator strength is significant has been shown.

This figure confirms, at least qualitatively, the experimentally observed redshift, and shows that the convergence towards the classical prediction is extremely slow. (For sodium, a Drude–Sommerfeld-type parametrisation of the sodium dielectric function leads to the value  $\omega_M = 3.49$  eV.) For small clusters ( $N < 100$ ), the absorption spectrum is broad and highly structured. This so-called fragmentation of the plasmon band is due to a phenomenon known as Landau damping and can be explained in a simple manner in terms of a discrete state coupled to a continuum, a problem dealt with in any standard textbook on quantum mechanics. The electron quantum state corresponding to the collective excitation of the conduction electrons has the same (or similar) energy as a multitude of excitations involving a single electron, referred to as particle–hole excitations in the language of nuclear or solid state physics. These one-electron transitions connect the occupied levels to the unoccupied

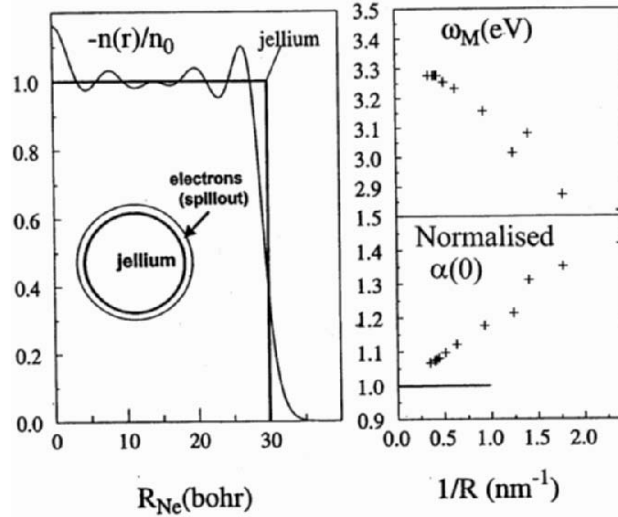




**Fig. 7.24.** Absorption cross-sections of sodium clusters  $\text{Na}_N$  calculated using the quantum TD LDA formalism within the framework of the jellium model. The spectra, highly structured at small cluster sizes, illustrate the fragmentation phenomenon, wherein the Mie resonance band, revealing collective excitation of the conduction electrons, splits into a number of separate peaks. This is induced by coupling with one-electron excitations. Also visible is the gradual convergence towards the classical prediction as the cluster size increases. The *dashed curve* in the *lower right-hand graph* is the classical result obtained with a Drude–Sommerfeld-type parametrisation of the sodium dielectric function ( $r_s = 3.93$  bohr,  $\Gamma = 0.1$  eV). The quantum finite size effects are responsible for the observed redshift (displacement of the resonance band towards low energies). The spectra comprising very narrow peaks for sizes  $N = 58, 138,$  and  $1284$  were calculated using a smoothing parameter ten times smaller (see Appendix L)

levels (hence above the Fermi level) in Figs. 7.11, 7.13 and 7.17, provided that the appropriate selection rules allow it.

The coupling between the excited collective state and one-electron excited states causes a quasi-Lorentzian broadening of the oscillator strength distribution connecting the electronic ground state and the excited collective state. For small sizes, the (pseudo-)continuum of one-electron excitations is in fact fundamentally discrete: the Landau damping phenomenon is reflected in this case by the presence of separate absorption peaks. When the cluster size increases, this fragmentation persists, but the oscillator strength is concentrated in an ever narrower spectral range around the plasmon resonance frequency, with a Lorentzian distribution. When  $N$  is very large, the calculated spectra,



**Fig. 7.25.** Illustration of the quantum effect known as spillout. Conduction electrons can appear beyond the classical radius  $R = R_{N_e} = r_s N_e^{1/3}$  of the cluster. Sodium clusters  $\text{Na}_N$  (monovalent  $N = N_e$ ,  $r_s = 3.93$  bohr, jellium model). *Left:*  $N = 440$ . The spillout of the electron density beyond the radius  $R$  makes a significant contribution to the finite size effects observed in the optical properties and static polarisation of metallic clusters. *Right:* Dependence of the surface plasmon angular frequency and static polarisability [normalised with respect to the classical value  $\alpha_{cl}(\omega = 0, N) = 4\pi\epsilon_0 R^3$ ] on the reciprocal of the radius  $R$ . Quantum finite size effects lead to a redshift of the resonance angular frequency  $\omega_M(N)$  relative to the classical prediction  $\omega_M^c = 3.49$  eV, and to an increase in the effective volume of the gas of conduction electrons

both classical and quantum, then become qualitatively similar, up to a redshift, and identical in the macroscopic limit (see Appendix L).

Let us now examine the observed redshift. In the classical theory, the electron density  $n(r)$  perfectly matches the positive ionic charge density  $n_+(r)$ . However, quantum calculations predict a significant overflow of electrons beyond the classical radius  $R_{N_e} = r_s N_e^{1/3}$  of the cluster, an effect known as spillout (see Fig. 7.25). This phenomenon cannot be neglected in the case of nanoscale clusters, where the Fermi wavelength  $\lambda_F$  is not negligible compared with the radius (the typical spatial scale of any significant variation in the density being of the order of  $\lambda_F$ ). Indeed, it has a great influence on the optical properties of metallic clusters. A simple classical argument shows that the spillout phenomenon is responsible for the experimentally observed redshift of the plasmon angular frequency  $\omega_M$ .

We use the jellium model for a cluster in vacuum. When  $r < R_{N_e}$ , the interaction between electron and jellium (potential energy) is equal to (see Sect. 7.4.1)

$$V_{\text{jel}}(r) = \frac{1}{4\pi\epsilon_0} \frac{qQ}{2R_{N_e}} \left[ \left( \frac{r}{R_{N_e}} \right)^2 - 3 \right] = \frac{1}{2} m(\omega_M^{\text{cl}})^2 r^2 - \frac{3}{8\pi\epsilon_0} \frac{qQ}{R_{N_e}}, \quad (7.107)$$

where  $\omega_M^{\text{cl}} = \omega_p/(3)^{1/2}$  is the classical value of the angular frequency of the Mie resonance. If the electron is inside the jellium ( $r < R_{N_e}$ ), it thus feels a restoring force  $\mathbf{F}$  such that

$$\mathbf{F} = -\nabla[V_{\text{jel}}(r)] = -m(\omega_M^{\text{cl}})^2 \mathbf{r}.$$

If the gas of conduction electrons is treated as a negatively charged rigid sphere, entirely contained within the radius  $r < R_{N_e}$ , the overall motion of this sphere is just that of a pure harmonic oscillator with angular eigenfrequency  $\omega_M^{\text{cl}}$ . Due to the spillout phenomenon, a significant fraction of the electrons feel a restoring force  $\mathbf{F}$  with smaller amplitude. It thus follows that

$$\mathbf{F} = -\frac{qQ\mathbf{r}}{4\pi\epsilon_0 r^3},$$

reducing the overall restoring force applied to the sphere and hence reducing the resonance frequency. As these outer electrons are less strongly bound to the centre of force, it also follows that the polarisation (and hence the displacement of this sphere relative to the jellium) induced by some static field will be smaller. This heuristic classical reasoning is in fact rigorously confirmed by the quantum analysis.

The effect of spillout can be quantified using the so-called sum rules obtained in the context of the jellium model. The sum rules are exact equations relating various moments  $M_k$  of the absorption cross-section, given by

$$M_k = \int \omega^k \sigma(\omega) d\omega, \quad (7.108)$$

to quantities characterising the electronic ground state of the cluster. For example,  $M_0$  is equal to

$$M_0 = \int_0^\infty \sigma(\omega) d\omega = \frac{\pi q^2}{2m\epsilon_0 c} N_e, \quad (7.109)$$

where  $N_e$  is the number of electrons. This rule, known as the Thomas–Reiche–Kuhn sum rule, is analogous to the rule applying to the hydrogen atom (one electron) which stipulates that the sum of the oscillator strengths between the ground state and excited states is equal to unity. We also have

$$M_2 = -\frac{q^2 \pi}{6\epsilon_0^2 c m^2} \int n_+(\mathbf{r}) n(\mathbf{r}) d^3\mathbf{r}. \quad (7.110)$$

Within the framework of the jellium model, the above integral is given by

$$\int n_+(\mathbf{r})n(\mathbf{r})d^3\mathbf{r} = -qn_0N_e \left(1 - \frac{\Delta N_e}{N_e}\right). \quad (7.111)$$

This follows straightforwardly since  $n_+(\mathbf{r})$  is either  $n_0$  for  $r < R_{N_e}$ , or 0 for  $r > R_{N_e}$ . The quantity  $\Delta N_e$  is the number of electrons (obviously, not necessarily an integer) located beyond the radius of the jellium. Assuming that all the oscillator strength, i.e.,  $\sigma(\omega)$ , is concentrated around the plasmon resonance  $\omega_M(N_e)$ , we have approximately

$$M_2 \approx [\omega_M(N_e)]^2 \int \sigma(\omega)d\omega = [\omega_M(N_e)]^2 M_0.$$

From (7.109), (7.110) and (7.111), we obtain the approximate relation

$$\omega_M(N_e) \approx \omega_M^{\text{cl}} \left(1 - \frac{\Delta N_e}{N_e}\right)^{1/2}. \quad (7.112)$$

The redshift of the resonance relative to the classical value  $\omega_M^{\text{cl}}$  thus constitutes a direct measure of the fraction of electrons overflowing from the jellium.

Furthermore, the static polarisability  $\alpha(0)$  is proportional to  $M_{-2}$  [see (7.108) for the definition]. Assuming that all the oscillator strength is concentrated around the plasmon resonance, and taking into account the fact that the asymptotic value of  $\alpha(0)$  ( $N_e$  large) must be the classical value  $4\pi\epsilon_0 R_{N_e}^3$ , it is easy to show that

$$\alpha(0) = 4\pi\epsilon_0 R_{N_e}^3 \left(1 - \frac{\Delta N_e}{N_e}\right)^{-1}. \quad (7.113)$$

Since the profile of the electron density  $n(r)$  is almost independent of the cluster size, we may deduce the following scaling law

$$\Delta N_e = -\frac{1}{q} \int_{R_{N_e}}^{\infty} 4\pi r^2 n(r) dr \approx aR_{N_e}^2 + a'R_{N_e} + s. \quad (7.114)$$

Keeping only the dominating term in this expansion (the surface term  $aR_{N_e}^2$ ), we obtain the following approximate relations:

$$\alpha(\omega = 0, R_{N_e}) = 4\pi\epsilon_0 (R_{N_e} + \delta)^3 \approx 4\pi\epsilon_0 R_{N_e}^3 \left(1 + \frac{3\delta}{R_{N_e}}\right) \quad (7.115)$$

and

$$\omega_M(R_{N_e}) \approx \omega_M^{\text{cl}} \left(1 - \frac{3\delta}{2R_{N_e}}\right). \quad (7.116)$$

These mean dependencies on  $1/R$  (see Fig. 7.25) thus directly reflect surface effects, and more specifically, the spillover of electrons beyond the classical

radius of the cluster, on the optical properties of clusters. Owing to the significant fragmentation of the plasmon band for small cluster sizes, the resonance angular frequency  $\omega_M$  has been defined in Fig. 7.25 as the angular frequency corresponding to the maximum of the smoothed absorption spectrum.

However, the spillout effect does not fully explain the experimentally observed redshift of the resonance frequency. Other ingredients need to be taken into account, in particular, the optical response of electrons in the ionic cores and pseudopotential effects.

The first ingredient arises partly from polarisation of ions (displacement of inner electron shells with respect to the nucleus) and partly from electronic transitions between completely filled inner bands, known as valence bands in solid state physics, and the unoccupied levels of the conduction band, situated above the Fermi level. Such transitions are referred to as interband transitions, as opposed to intraband transitions between occupied and unoccupied levels of the conduction band. These polarisation and absorption effects, which are relatively unimportant in the alkali elements, will be illustrated later when we discuss the optical properties of the noble metals.

As we saw in the Sect. 7.4.2 when discussing electronic supershells, the profile of the effective potential  $V_{\text{eff}}(r)$ , and hence also the electron density  $n(r)$ , is softened when the electron-ion interaction pseudopotentials  $v_{\text{ps}}(\mathbf{r} - \mathbf{R}_i)$  are taken into account. This softening increases the electron spillout phenomenon relative to the predictions of the jellium model. In addition, the non-local character of the pseudopotentials has the consequence, at least formally, of ‘increasing’ the mass of the electron. (We speak of the effective electron mass  $m_{\text{eff}}$ .) This leads to a reduction in the bulk plasma angular frequency  $\omega_p$ , and hence also a reduction in  $\omega_M$ . Note that this property, which expresses the fact that the conduction electrons do not really constitute a free electron gas, is not a surface effect, but is rather a characteristic of the bulk phase, e.g.,  $m_{\text{eff}}(\text{Na}) \approx m$  and  $m_{\text{eff}}(\text{Li}) \approx 1.4m$  (sodium is thus the archetypal ‘simple’ metal).

In the case of the noble metals, the optical properties are more difficult to interpret because the closed valence bands ( $3d^{10}$ ,  $4d^{10}$  and  $5d^{10}$  for copper, silver and gold, respectively) are close in energy to the Fermi level  $E_F$ , the interband thresholds  $\hbar\Omega_{\text{ib}}$  ( $\hbar\Omega_{\text{ib}}$  is the energy difference between  $E_F$  and the top of the  $d$  band) being of the order of 2 eV (Cu), 2 eV (Au) and 4 eV (Ag), respectively. Interband excitations between levels of the  $d$  band and the (unoccupied) levels of the conduction band are thus superposed upon the intraband electronic excitations from the  $sp$  conduction band. Naturally, the optical response does not reduce to a simple superposition of these two types of transition, because the electrons are not independent, but are in fact coupled by the Coulomb interaction, as demonstrated by the collective Mie excitation, and also because of mutual screening effects. Equation (7.104) which gives approximately the angular frequency of the Mie resonance, clearly illustrates this, since this angular frequency is significantly shifted towards low energies due to the large values of  $\varepsilon_1^d(\omega_M)$  in these metals. For example,

a simple Drude–Sommerfeld-type parametrisation of the dielectric functions of silver or gold [ $r_s(\text{Ag}) \approx r_s(\text{Au}) \approx 3$  bohr,  $m_{\text{eff}}(\text{Ag}) \approx m_{\text{eff}}(\text{Au}) \approx m$ ] leads to the value  $\omega_M \approx 5.2$  eV for clusters in vacuum ( $\epsilon_m = 1$ ). Taking into account the complex dielectric function  $\epsilon^d(\omega)$  associated with core electrons, one obtains the values  $\omega_M(\text{Ag}) \approx 3.5$  eV and  $\omega_M(\text{Au}) \approx 2.5$  eV, very close to the experimentally measured values.

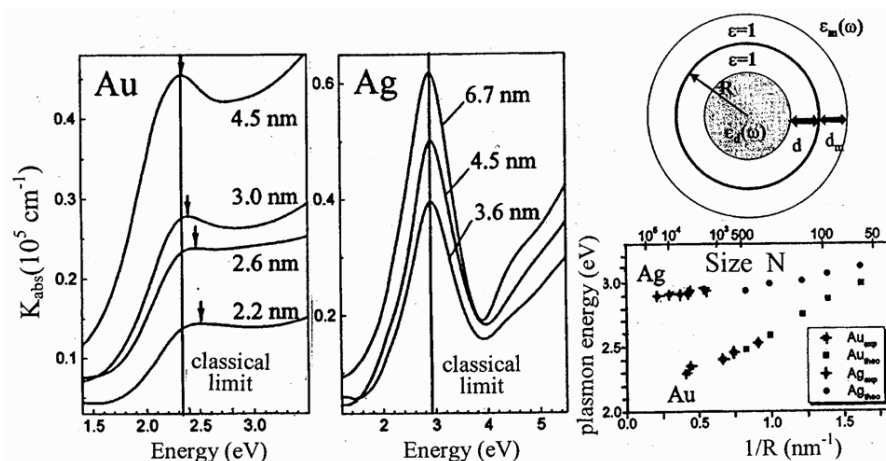
Figure 7.26 shows the absorption spectra of composite thin films made from gold or silver clusters embedded at low densities (to avoid dipole interactions between clusters) in a porous alumina matrix. The interband transitions are clearly visible in the figure. The surface plasmon band for gold is considerably broadened by the fact that the resonance angular frequency is situated above the interband threshold  $\Omega_{\text{ib}}$  ( $\hbar\Omega_{\text{ib}} \approx 2$  eV): added to the ‘intra-band’ fragmentation, there is also a broadening caused by coupling between the discrete state corresponding to the collective excitation and the continuum of interband excitations. It should be stressed that the broadening observed in Fig. 7.26 is to a large extent inhomogeneous, especially for silver, due to the broad distributions of size, shape and local environment characterising composite films. (The term ‘inhomogeneous’ means that the broadening is not intrinsic to an individual cluster, but results from averaging over a large number of clusters.)

Finite size effects are much smaller than those measured on clusters of alkali metals, although the smaller Wigner–Seitz radii  $r_s$  of the noble metals would suggest that the influence of the surface would be more pronounced for fixed size  $N_e$ . Further analysis shows that:

- In gold clusters, the Mie resonance shifts significantly towards the blue in a way that goes roughly as  $1/R$  (a blueshift therefore, rather than the redshift observed for the alkali elements). It also broadens and fades out as the size decreases.
- Finite size effects are almost non-existent in silver clusters.

The blueshift observed for gold is barely visible in Fig. 7.26. In fact, the damping and increased broadening due to fragmentation, which are directly correlated to the density of the interband excitations with energy close to  $\omega_M$ , are an unmistakable signature of this blueshift. As the energy increases above the interband threshold  $\hbar\Omega_{\text{ib}}$ , more and more occupied states in the  $d$  band and unoccupied states in the  $sp$  conduction band can become involved. Then  $\epsilon_2^d(\omega_M)$  increases abruptly above  $\Omega_{\text{ib}}$ .

In order to interpret the experimental results, dielectric effects due to the ionic cores and the matrix have been incorporated into a jellium-type model (homogeneous ion distribution) and the optical response calculated using the TD LDA formalism (internal fields and optical responses of the various dielectric media are calculated self-consistently). A further ingredient which has been well established by the work of many research groups in the context of metallic surfaces and correlated with the localisation of atomic  $d$ -orbitals relative to the orbital of the outer  $s$  electron is the existence of a



**Fig. 7.26.** Absorption cross-sections of gold clusters  $Au_N$  and silver clusters  $Ag_N$  embedded in a porous alumina matrix with  $\epsilon_m \approx 2.7$  in the visible, for various samples (thin composite films with thicknesses of the order of a hundred nanometers). The mean diameters characterising the size histograms determined by transmission electron microscopy are indicated. Inhomogeneous effects (distributions in size, shape and local environment) are largely responsible for the broadening of the plasmon resonance band, especially in the case of the silver clusters. *Top right:* Model used to interpret results. Relative dielectric functions indicated in the diagram are those associated with media other than the conduction electron gas. The model, which is of jellium type, takes into account the dielectric properties of the matrix and ionic cores in a self-consistent manner. It also includes an inner ring of reduced ionic polarisability [ $\chi_d(\omega) = 0$ ], and a ‘vacuum ring’ simulating the local porosity at the interface. *Bottom right:* Comparison between the dependencies of the observed and calculated resonance frequencies on the reciprocal of the radius [25]

thin surface layer where the polarisation of the ionic cores is ineffective. The internal dielectric medium is thus limited to a sphere of radius  $R_{N_e} - d$ . The influence of this surface zone can be immediately analysed using the equation (7.104) giving the Mie angular frequency, by considering the following limiting cases:

- $d = R_{N_e}$  [small clusters,  $\epsilon^d(\omega) = 1$ ],
- $d/R_{N_e} \approx 0$  [large clusters,  $\epsilon^d(\omega) \gg 1$  throughout almost the whole volume of the cluster].

The existence of this zero polarisation layer will thus induce a blueshift in the resonance frequency, and the blueshift will increase as the cluster size decreases.

Another crucial factor confirmed by much experimental work is the influence of the interface, and in particular, the local porosity of the matrix at the interface (intrinsic porosity of the matrix, contact defects between cluster and matrix, roughness of the cluster surface). A thin vacuum layer of thickness

$d_m$  has also been added to the model to simulate this reduced local polarisation. A similar analysis to the one above shows that this effect also leads to a blueshift.

It should be emphasised that the thicknesses  $d$  and  $d_m$  are very small, being of the order of one atomic radius. The values of these two effective parameters, which are difficult to estimate a priori in a non-granular model of jellium type, have been fitted independently in a first stage, by comparison with the experimental results obtained using silver clusters, either in free phase or in the solid phase (spectra in Fig. 7.26). As the relevant physical quantities are similar for gold and silver [same  $r_s$  and effective masses, similar radial extent of the  $4d$  orbital (Ag) and  $5d$  orbital (Au)], and since the composite films are produced under identical conditions, the same values of  $d$  and  $d_m$  should apply to both elements. Figure 7.26 (bottom right), which shows the  $1/R$  dependence of the resonance energy  $\hbar\omega_M$  for the two noble metals, confirms the validity of this model.

Finite size effects in gold and silver clusters thus result from competition between two opposing effects: on the one hand, electron spillout which induces a redshift, and on the other, reduced polarisability layers which induce a blueshift. These two effects cancel almost exactly for silver. Since the two metals have the same value of  $r_s$  (to within a few  $10^{-3}$ ), the spillout phenomenon is identical for silver and gold. But since the real component  $\varepsilon_1^d(\omega)$  is bigger for gold than for silver [because  $\Omega_{ib}(\text{Au}) \ll \Omega_{ib}(\text{Ag})$ ], the blueshift will dominate slightly in gold clusters [see (7.104)]. Another, more subtle effect which helps to decide the outcome of this competition is the strong spectral dependence of  $\varepsilon_1^d(\omega)$  near the resonance frequency.

Throughout this section, we have limited the discussion to the optical properties of an isolated homogeneous spherical metal cluster in the dipole approximation. These properties can in fact be varied in an infinite number of combinations by considering more complex systems. A deformation of the cluster gives rise to several resonance bands, the restoring force of the electron gas depending on the direction of oscillation induced by the field (with three resonance frequencies for the most general ellipsoid). Composite clusters containing several different elements, in particular, clusters with a core-shell configuration (one material surrounding the other) for immiscible systems (metal-metal or metal-insulator), have specific optical properties that can be modified by adjusting the proportions of the constituent elements.

For these systems, quantum finite size effects are also important, because new characteristic length scales, e.g., layer thicknesses, dimensions of component entities, must be compared with the Fermi wavelength  $\lambda_F$ . It follows that the ‘chemical’ interfaces (the interfaces between two different materials) do not necessarily give rise to classical interface modes, from the point of view of optical absorption. For example, in clusters formed by alternating concentric layers of two simple metals with very different densities, the quantum spectrum calculated using the jellium model only features a single resonance



band, similar to the one obtained with a homogeneous cluster of the same mean density.

Finally, when there is a high density of clusters in the composite film, the dipole interactions between clusters and large scale multipole effects must be taken into account, i.e., the full Mie theory, not limited to the dipole term, must be used. This will confer new optical properties upon such materials.

## 7.5 Preparation Methods

There are far too many methods for preparing clusters to be able to present them all here. We shall therefore limit the discussion to the so-called bottom-up approach, wherein clusters are assembled from their individual elements. The two channels used are the physical channel and the chemical channel. In the former, clusters are produced in the ‘pure’ state, whereas in the latter, clusters are surrounded by ligands formed from molecules with differing degrees of complexity, based on carbon, nitrogen, oxygen, hydrogen, and so on.

### 7.5.1 Gas Phase Physical Methods

We shall only discuss here the production of clusters via isentropic expansion. However, the nucleation processes described later are perfectly universal, whatever the preparation method.

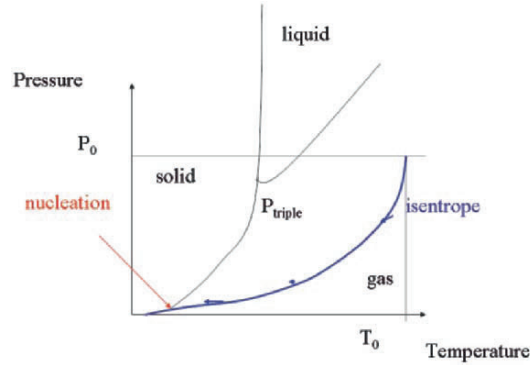
#### Basic Idea

Any element of the periodic table is characterised by its phase diagram  $P(T)$ , where  $P$  is the pressure and  $T$  the temperature. If we prepare the element that we wish to obtain in cluster form from the vapour phase by heating it in thermodynamic equilibrium, there will be no nucleation. In order to obtain nucleation, the system must undergo a gas–liquid (GL) or gas–solid (GS) transition (see Fig. 7.27).

This happens if the system moves along an isentrope (constant entropy) which crosses the equilibrium diagram. It can be shown that, for an isentropic expansion through a nozzle, the following relation holds between the temperature  $T$  and the pressure  $P$ :

$$\frac{P}{P_0} = \left( \frac{T}{T_0} \right)^{\gamma/\gamma-1}, \quad (7.117)$$

where  $P_0$  and  $T_0$  are constants, and  $\gamma$  is the ratio of specific heats, i.e., 1.66 for a perfect monatomic gas. The curve in the  $P(T)$  diagram will thus have a significant slope. When it crosses the GL or GS equilibrium curve, there will



**Fig. 7.27.** Phase diagram  $(P, T)$  and isentrope given by (7.117). The movement of the system along the isentrope BC is shown by *arrows*. Nucleation occurs when the isentrope crosses the gas–solid curve at point X. Depending on the element, the isentrope may also cross the gas–liquid curve

be a phase transition and nucleation. This nucleation has to be stopped in order to limit the final size of the clusters.

The isentropic expansion is obtained using supersonic beams. To this end the initial gas expands in the vacuum via a nozzle with diameter much smaller than the mean free path of the gas molecules. This requires a very high saturated vapour pressure, of the order of one atmosphere or more. A quick calculation shows that this kind of expansion is impractical for refractory materials like carbon. Such an element would require initial temperatures above 5000 K! To solve this problem, a second gas is used as a thermostat (the seeded beam), as we shall now describe.

## Homogeneous Nucleation

### *Classical Model*

Nucleation refers to the homogeneous appearance of germs. Suppose that during the sudden cooling of the vapour during expansion, spherical droplets of radius  $R$  are observed to form. The free enthalpy of formation of a droplet comprises two terms: a term related to the volume and a term representing the contribution of the surface (surface energy):

$$\Delta G(R, T, P) = \Delta G_{\text{vol}} + \Delta G_{\text{surf}} . \quad (7.118)$$

If the stable phase is the solid phase,  $\Delta G_{\text{vol}}$  is negative and energy is gained by forming the drop. The surface contribution, on the other hand, will be positive. If we define the free enthalpies per unit volume  $\Delta g_{\text{vol}}$  and per unit surface  $\Delta g_{\text{surf}}$ , we obtain

$$\Delta G_{\text{vol}} = -\frac{4\pi}{3} R^3 \Delta g_{\text{vol}} \quad (7.119)$$

and

$$\Delta G_{\text{surf}} = 4\pi R^2 \gamma = 4\pi R^2 \Delta g_{\text{surf}} , \quad (7.120)$$

where  $\gamma$  is the surface tension. Hence,

$$\Delta G(R, T, P) = -\frac{4\pi}{3} R^3 \Delta g_{\text{vol}} + 4\pi R^2 \Delta g_{\text{surf}} , \quad (7.121)$$

and this energy is minimised if

$$\frac{\delta \Delta G(R, T, P)}{\delta R} = 0 . \quad (7.122)$$

Therefore the critical radius  $R^*$  above which nucleation will occur is given by

$$R^* = \frac{2\gamma}{\Delta g_{\text{vol}}} . \quad (7.123)$$

$\Delta g_{\text{vol}}$  is homogeneous with one pressure  $P = (\delta G / \delta V)_T$ . The last equation is just the Gibbs equation

$$\Delta g_{\text{vol}} = \Delta P = \frac{2\gamma}{R^*} . \quad (7.124)$$

Finally, we obtain

$$\Delta_{\text{min}} G(R, T, P) = \frac{16\pi\gamma^3}{3\Delta g_{\text{vol}}} . \quad (7.125)$$

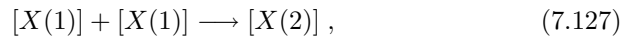
In a classical way, we may define a nucleation rate  $P_R$ , the probability of having a droplet of radius  $R$ , on the basis of an Arrhenius law

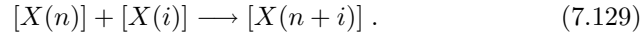
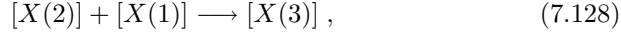
$$P_R \propto \exp \frac{-\Delta G_{\text{min}}(R, T, P)}{k_B T} . \quad (7.126)$$

Once beyond the critical radius, if the temperature is low enough, homogeneous nucleation is a fast process. The critical radius depends on the characteristics of the element. We shall investigate its physical origin in more detail.

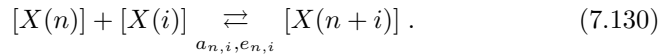
### *Statistical Model*

During the expansion, there are collisions between the various constituents of the vapour. To begin with, we shall consider just one constituent  $X$ . The droplet contains  $n$  atoms and has concentration  $[X(n)]$ . When nucleation is just beginning, there will be collisions between two monomers  $X(1)$ . Such a collision can form a dimer  $X(2)$ , and this may collide with a monomer or another dimer. Quite generally, the following reactions can occur:

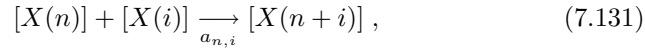




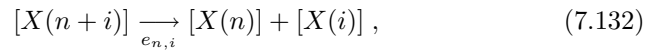
The rate of production of  $[X(n+i)]$  depends on two terms: the probability of forming new complexes  $[X(n+i)]$  and the probability of destroying existing complexes  $[X(n+i)]$ . Each reaction results from two simultaneous reactions:



Let us examine each reaction:



where  $a_{n,i}$  expresses the probability of forming a complex  $X(n+i)$ , and

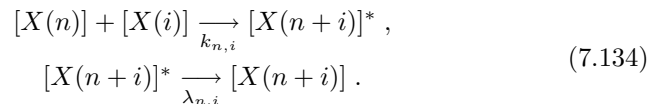


where  $e_{n,i}$  expresses the probability of destroying the complex. A complete solution can be obtained from the kinetic equations, which yield  $N$  coupled differential equations:

$$\begin{aligned} \frac{d[X(n)]}{dt} = & \sum_i a_{n-i,i} [X(n-i)][X(i)] + \sum_i e_{n,i} [X(n+i)] \\ & - \sum_i e_{n-i,i} [X(n)] - \sum_i a_{n,i} [X(n)][X(i)] . \end{aligned} \quad (7.133)$$

The first term on the right-hand side represents the probability of forming the complex  $X(n)$ . The second such term represents the probability of destroying a complex  $X(n+i)$  with fragment  $X(n)$ . The third term represents the probability of destroying the complex  $X(n)$  (hence the minus sign), and the fourth term represents the probability of obtaining a bigger complex  $X(n+i)$ . This system cannot be solved explicitly. We shall assume that the reaction forming a complex takes place in two consecutive stages: first the formation of a quasi-complex given uniquely by the collision probability, hence by the concentration of each constituent; and second, the probability that this quasi-complex survives long enough to take part in a further reaction. The latter probability depends on the binding energy of the atoms making up the complex and on the temperature of the complex.

With this hypothesis, (7.130) reduces to two completely independent reactions, while  $e_{n,i}$  and  $a_{n,i}$  are coupled:



The first term expresses the probability of a collision between two species  $X(n)$  and  $X(i)$  and can be calculated by means of the kinetic theory of gases using only the collision frequency related to the partial pressure. The second term depends on a sticking coefficient and expresses the probability that the complex formed by collision between  $X(n)$  and  $X(i)$  manages to survive. The asterisk indicates that the complex is in an excited state. Indeed, when two species condense to form a larger complex, the latter must absorb the condensation energy, mainly in the form of vibrational energy. Raising the temperature increases the probability of fission, and all the more so as the complex is small. The term  $\lambda_{n,i}$  is called the sticking coefficient. Physically, it expresses the probability of stabilising the complex  $[X(n+i)]$  when the excited complex  $[X(n+i)]^*$  has been formed by collision. If the sticking coefficient is equal to unity, the complex will always be stable and will survive. However, if the coefficient is zero, there is no chance of obtaining a stable complex. In this case, the complex  $[X(n+i)]^*$  is destined to destroy itself by fragmentation if  $[X(i)]_i \geq 2$ , or by evaporation if  $[X(i)]_i = 1$ . Since the value of  $\lambda_{n,i}$  ranges between zero and unity, we may define an empirical equation satisfying this constraint:

$$\lambda_{n+i} = 1 - \exp\left(-\frac{\tau_{v,n+i}}{\tau_{s,n+i}}\right), \quad (7.135)$$

where  $\tau_{s,n+i}$  is the time required for the complex  $X_{n+i}$  to collide with an atom or another complex. This time is entirely predetermined by the initial cluster density.

$\tau_{v,n+i}$  is the lifetime of a quasi-complex whose temperature after the nucleating collision  $X_n + X_i$  is  $k_B T$ . According to the RRK model [26], this term is given by

$$\tau_{v,n+i} = \nu^{-1} \left( \frac{E_{n+i}}{E_{n+i} - D_l} \right)^{3N-7}, \quad (7.136)$$

where  $E_{n+i}$  is the total energy stored in the complex,  $\nu$  is the characteristic vibration frequency,  $N$  is the number of atoms ( $N = n + i$ ),  $D_l$  is the dissociation energy of a mode  $l$  representing the least strongly bound mode of the complex. In other words, the complex will break at the point where the atoms are least strongly bound. In the liquid-drop model, where all the bulk modes  $l$  are identical, dissociation will tend to occur at the surface. This mechanism corresponds to evaporation of atoms or fragments from the surface when a crystal is heated. If we assume that  $D_l$  varies only slightly with the cluster size (see this problem in Sect. 7.1.1), the time  $\tau_{v,n+i}$  will nevertheless vary considerably with  $N$  (power law).

This brings us to the idea of a critical germ  $r^*$ . This is the minimal size of a complex such that the increase in temperature caused by condensation of an atom or cluster does not lead to fission of this complex of radius  $r^*$  and containing  $N$  atoms. We may apply the statistical criterion

$$\lambda_{n+i} = 1 - \exp(-1), \quad (7.137)$$

whence, according to (7.135),

$$\tau_{v,n+i} = \tau_{s,n+i}. \quad (7.138)$$

As an example, consider a tetramer with binding energy around 3 eV (a typical value for a metallic cluster). We obtain  $\tau_{v,n+i} \sim 10^{-8}$  s. In contrast, for a cluster with a low binding energy ( $\sim k_B T$ ) of van der Waals type, this time is of the order of  $10^{-12}$  s. The time  $\tau_{s,n+i}$  depends on the pressure and, in standard pressure conditions, viz.,  $P > 10^5$  Pa, may reach values below  $10^{-8}$  s. If the quasi-complex has the shape of a droplet of radius  $R^*$ , we have the geometric relation

$$R^* = N r_a, \quad (7.139)$$

where  $r_a$  is the Wigner–Seitz radius of the atom. Finally, using (7.135) and (7.137),

$$R^* = r_a \frac{1}{3} \left\{ \frac{\ln[E_{n+i}/(E_{n+i} - D_l)]}{\ln(\nu \tau_{s,n+i})} \right\}^{1/3}. \quad (7.140)$$

For elements with strong binding energies, e.g., metals, covalent elements, it is found that the critical germ is close to the dimer  $R_* \sim r_a$ . This validates the use of statistical models which more accurately describe the first stages of nucleation, rather than the liquid-drop model which works for macroscopic systems.

#### *Practical Details*

Using isentropic expansion of a vapour containing a single element X, it is found that the time  $\tau_{s,n+i}$  is relatively long and does not lead to a good sticking coefficient. To solve this problem, a noble gas is introduced to function as a heat bath. With helium, the very low value of  $D_l$  (van der Waals binding) is indeed too low to allow nucleation of helium. However, via the van der Waals force, the complex  $X_{n+i}$  can relax by collision with atoms of the noble gas. If the pressure of the noble gas is very high compared with the partial pressure of the species X, the time  $\tau_{s,n+i}$  will be governed by the helium pressure. One thus uses a mixture of a noble gas and the vapour of the element to be nucleated. The vapour can be obtained in different ways: direct heating of the element by the Joule effect, laser ablation, ion bombardment, plasma reaction, and so on.

## **7.5.2 Liquid Phase Chemical Methods**

### **Metal Colloids**

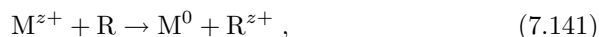
#### *Synthesis by Reduction of the Metal Salt*

Any metal species with high enough standard reduction potential can be synthesised by chemical reduction of the corresponding metal ion, in such a way

**Table 7.4.** Standard reduction potentials in aqueous solution at 25°C with respect to the normal hydrogen electrode (NHE)

Couple	Potential [V vs. NHE]
Ag <sup>+</sup> /Ag	0.80
Au <sup>+</sup> /Au	1.68
Au <sup>3+</sup> /Au <sup>+</sup>	1.29
Pd <sup>2+</sup> /Pd	0.83
Pt <sup>2+</sup> /Pt	1.2
Cu <sup>2+</sup> /Cu	0.34
Cu <sup>+</sup> /Cu	0.52
Cu <sup>2+</sup> /Cu <sup>+</sup>	0.16
Ni <sup>2+</sup> /Ni	-0.23
Zn <sup>2+</sup> /Zn	-0.76
Na <sup>+</sup> /Na	-2.71
Li <sup>+</sup> /Li	-3.04
K <sup>+</sup> /K	-2.92

that the reduced species is stable and will not immediately reoxidise. This reaction, of type



involves a metal ion  $M^{z+}$  reduced to the form  $M^0$  and a reducing agent  $R$ . It is favoured thermodynamically if

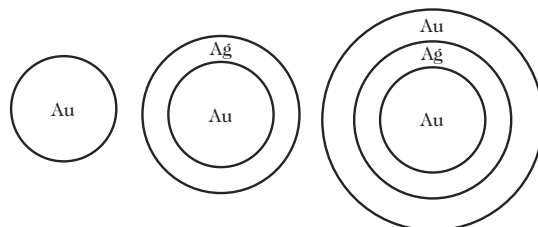
$$E^0(M^{z+}/M) - E^0(R^{z+}/R) > 0, \quad (7.142)$$

where  $E^0(M^{z+}/M)$  and  $E^0(R^{z+}/R)$  are the standard reduction potentials of the pairs  $M^{z+}/M$  and  $R^{z+}/R$ . The main candidates are thus the noble metals, such as gold, silver and platinum, whose standard reduction potentials are given in Table 7.4 relative to the standard hydrogen electrode.

In contrast, metals like nickel, zinc or the alkali metals cannot be obtained in colloid form by this route, because the backward reaction is often thermodynamically favoured due to the species present in the solution, especially in the aqueous phase. One may turn to organic solvents to avoid this difficulty, or use a physical route in the gas phase, e.g., laser vaporisation.

There is a certain freedom in the choice of reducing agent. However, this choice does determine the parameters of the reduction reaction, in particular, the reaction rate constant and the exact reaction mechanism. Indeed, citrate is a reducing agent commonly used to prepare gold clusters, and it leads to a mechanism involving the formation of complexes between gold and citrate or an intermediate species. Seeds can be introduced into the solution at this stage, on which the clusters will proceed to grow after this induction period.

Apart from citrate, another commonly used reducing agent is sodium borohydride  $\text{NaBH}_4$ . This is a highly efficient reducing agent, leading to very small



**Fig. 7.28.** Morphology of spherical clusters. Spherical core-shell and multilayer core-shell Au–Ag clusters

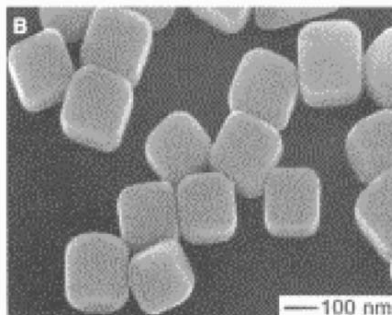
clusters of the order of 3.5 nm in diameter. This chemical reduction is achieved at a temperature close to 370 K, in order to obtain fast enough reaction kinetics in relation to growth termination processes, such as surfactant adsorption. For  $\text{NaBH}_4$ , the reaction can simply be carried out at room temperature. It can be introduced either at the end of the reaction or at the very beginning. Citrate also serves to enhance stabilisation if introduced in excess. Heating is traditionally achieved by flame or water bath, but other methods can be used, such as heating by infrared electromagnetic radiation.

The clusters obtained in this way then serve as seeds for growing larger clusters. This last stage is achieved using a moderate reducing agent such as hydroxylamine or ascorbic acid in the presence of a surfactant or capping agent, i.e., an organic compound adsorbing easily onto the surface of the metallic cluster. Surfactants such as trimethylammonium bromide or sodium dodecylsulfate thus allow competition between the growth reaction and the termination reaction by adsorption. The final diameter of the clusters is controlled by the exact stoichiometry of the reagents and can vary between 3.5 nm and more than 100 nm.

The synthesis of clusters of alloys involving two metals which have very similar lattice parameters, such as silver and gold, is achieved by modifying the composition of the initial metal salt. For these two metals, in particular, the initial composition is recovered in the final composition of the alloyed metal clusters. For other metals, the initial composition is not necessarily conserved in the final composition.

Finally, by growing a metal shell on an initial core during a second stage of the synthesis reaction, heterogeneous morphologies can be obtained. In particular, metastable clusters can be synthesised, with non-equilibrium morphology. This is the case for clusters with a gold core and a platinum shell, for which the stable form comprises a platinum core covered by a gold shell, or gold–silver multilayer systems (see Fig. 7.28). The thermodynamically stable morphology can then be recovered by annealing at high temperature. As an example, the practical details for synthesis of gold clusters are given in Appendix M.





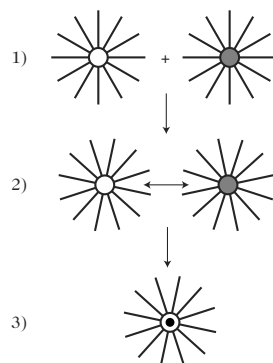
**Fig. 7.29.** Scanning electron microscope image of cubic silver clusters [27]

Chemical methods of synthesis lead to faceted clusters with several crystal planes accessible to the surrounding liquid phase. By introducing organic compounds into the reaction medium, capable of adsorbing onto the metal surface, growth can be pursued in certain privileged directions. Indeed, the free energy of adsorption differs from one crystal plane to another, and growth in certain planes is thus blocked by adsorption of a complete layer of adsorbates, deactivating these sites with regard to further growth.

These more sophisticated methods are important, because they give access to shapes other than the sphere, such as cubes and prisms (see Fig. 7.29). Such techniques are currently receiving a great deal of attention as a way of diversifying the morphological characteristics of clusters.

#### *Reduction of the Metal Salt by Irradiation*

In the various methods using irradiation, the reducing agent is itself the product of a reaction initiated by absorption of ionising particles, either electrons with kinetic energies of a few MeV, or gamma, X or visible photons. In all these cases, the reaction is carried out without an initially present chemical reducing agent. Indeed, metal ions are reduced by reducing species which are formed by photolysis of the water in the aqueous liquid phase or activation of species initially present in solution. The absorption of a gamma or X photon, or simultaneous absorption of several visible photons, photodissociates the water molecules into hydrogen or hydroxyl radicals, or produces solvated electrons in the solution. These chemical species are highly reactive and easily reduce the metal ions present in the solution. The extent of the reaction is no longer controlled by the relative concentrations of the initial reactive species, but rather by the duration, the flux and the energy of the ionising particle bombardment. A termination reaction effected by surfactant adsorption is also useful in this scenario, to provide better control over the size and shape distribution of the clusters.

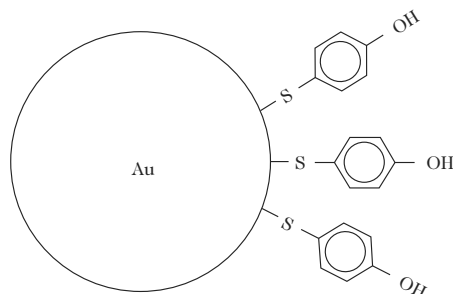


**Fig. 7.30.** Reverse micelle reaction. (1) Preparation of two independent solutions of reverse micelles, e.g., NaAOT and AgAOT. (2) Mixing the two solutions with exchange of micellar media. (3) Reduction of metal ions and growth of clusters inside the micelles

#### *Reduction by Reverse Micelles*

Reduction in the presence of reverse micelles, as illustrated in Fig. 7.30, is identical to the reduction process in solution, except that the reaction volume is much smaller and is defined by the micelle. This micelle is in fact formed by phase separation of organic compounds that are insoluble in the solvent. For this synthesis, the micelles are generally made from sodium diethyl sulfosuccinate, also known as NaAOT, dispersed in hexane or isooctane in the presence of water, at a concentration above the critical concentration above which micelles form. The solution is then formed by micelles containing a volume of water defined by the various concentrations of NaAOT and the organic and aqueous solvents.

The next step is to mix two solutions of reverse micelles, one containing silver metal ions introduced in the form of silver diethyl sulfosuccinate in a partial substitution for NaAOT, and the other containing a reducing agent such as hydrazine. When the two solutions are homogenised, the micellar media undergo exchange, thereby allowing the reducing agent to come into contact with the metal ions. The last stage of this procedure requires the introduction of an agent capable of terminating the reaction and flocculating the solution. A commonly used agent for this purpose is dodecanethiol. The solution is then filtered and redispersed in hexane several times in order to remove the excess dodecanethiol that has not been adsorbed. Finally, the solution must sometimes be centrifuged to separate clusters of different sizes and free organic species. The cluster size distribution before centrifugation may be as high as 30–40% and this last stage is often repeated several times to reduce this to 10–15%. This is a versatile method which can be extended to other chemical species to prepare cobalt or silver sulfide clusters, for example.



**Fig. 7.31.** Gold clusters functionalised by a thiophenol layer. The OH bond can be broken for subsequent grafting of functional groups such as acid or amine functions

#### *Cluster Stabilisation and Derivatisation*

These different methods lead to solutions of clusters and the main concern of the user is the stability of the solution over time. Apart from problems related to the reactivity of the clusters and the formation of a surface oxide layer, for example, there is also the problem of the coalescence of particles as the solution ages.

This process results from competition between the attractive van der Waals forces between the clusters and repulsive forces of electrostatic or steric origin. Metallic clusters in particular are highly polarisable, so that strong induced dipole/induced dipole forces arise at short distances, leading to coalescence of the particles into large entities that may sediment out. In order to counter these attractive forces, the usual approach is to incorporate some stabilising agent. For example, when clusters are synthesised by reduction with citrate, a large excess of citrate is used, so that the resulting metallic clusters are coated with a layer of adsorbed citrate. The clusters are thus negatively charged and the repulsive electrostatic force between them opposes the attractive van der Waals force. The citrate is thus a reducing agent and a stabilising agent at the same time.

When using other methods, such as irradiation or reduction in reverse micelles, stabilisation is achieved by a species adsorbed at the surface of the cluster. This species may be neutral or ionised. In the latter case, stabilisation is electrostatic. In the former, it is due to steric hindrance of the adsorbed compounds, which prevents coalescence of particles into larger entities.

The idea of stabilising clusters is sometimes just a first step towards the design of clusters with their own function, e.g., for biological compatibility or for molecular recognition. In this case, the species fixed at the surface can be a starting point for subsequent chemical syntheses, whereby precise chemical functions are grafted onto the cluster surface, such as amine or carboxylic acid functions, or whole molecular compounds such as fluorophores and more recently, DNA sequences. The latter developments are currently attracting much attention for the synthesis of metallic clusters.

## 7.6 Cluster or Colloid Assemblies

Over the last few decades, thin film materials produced by physical or chemical methods have been greatly developed due to the large number of applications in a wide range of different fields: design of components and circuits in microelectronics, sensor technology, and surface coatings to protect against corrosion, wear, light reflection (anti-reflective layers), and so on. Depending on the application, thin films of different types (metallic, semiconductor, transparent insulating oxides, etc.) and different thicknesses (from a few nanometers to a few micrometers) are used. The areas to be coated also vary, and can be very large in certain cases, e.g., protective or optical coatings. In all these applications, specially designed atomic and molecular deposition methods have been developed to carry out controlled production of thin film materials, and equipment is commercially available for use both in research laboratories and industry.

Among the most commonly used methods, we may cite physical methods wherein gas phase atomic or molecular beams are produced to carry out deposition onto the appropriate surfaces, and chemical methods which generally operate in the liquid phase, such as electrodeposition or sol-gel processes. Depending on the technique and deposition conditions, the structures of the thin film materials thereby obtained can be extremely varied, since one may produce amorphous, polycrystalline or crystalline phases.

In this general context, the idea of producing thin films using clusters or colloids rather than atoms or molecules was put forward in the 1980s, when the novel properties and structures of these nanoscale objects first came to light. These features are direct consequences of confinement effects and the high surface-to-volume ratio, which are in turn both related to the small size of the objects. Hence, if it was possible to deposit clusters or colloids on a surface without destroying them in the process, in order to grow thin films by stacking up these basic building blocks, one could hope to obtain a material with a nanostructured morphology, somewhere between that of amorphous and polycrystalline materials, which would conserve in memory the novel properties of the basic building blocks.

A new field thus opened up in the synthesis of functional nanostructured materials, offering interesting prospects for applications in key areas such as nanoelectronics, nanooptics, nanomagnetism, nanobiology, and so on. In this section, we shall describe methods commonly used to prepare nanostructures and nanostructured thin film materials by assembling clusters previously formed in the gas phase or colloidal clusters prepared in the liquid phase. The nucleation mechanisms and growth characteristics making it possible to explain and control the nanostructured morphologies of these systems are discussed along with examples of these types of material.

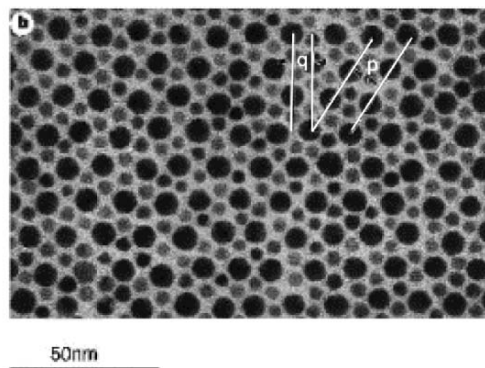
### 7.6.1 Assemblies of Metallic Clusters

The aim when metallic clusters are condensed into two- and three-dimensional objects is to build large organised entities with novel properties. These entities may be either pure or hybrid, i.e., containing both clusters and organic compounds. In solution, 3D assemblies of metallic clusters are easily obtained by chemical synthesis in solution. Indeed, if we take the example of gold clusters obtained by reduction from citrate, the clusters are stabilised by electrostatic forces due to the presence of an excess of citrate adsorbed on the cluster surfaces. Introducing an organic compound with high affinity for metallic gold, e.g., pyridine, the adsorbed citrate is replaced by the electrically neutral pyridine and the cluster solution thereby destabilised. A process of coalescence is thus triggered, initially between the particles and the first small clusters, then between the clusters themselves. This process, whose reaction rate can be controlled via the concentration of pyridine introduced into the solution, thus leads to the formation of large 3D assemblies with geometry described by the theory of fractals.

Hybrid assemblies containing organic compounds with specific optical properties, e.g., Rhodamine 6G molecules with their well known fluorescence properties, are straightforwardly produced by replacing pyridine by the chosen compound. The drawback with this method in solution is that it leads to assemblies with no well-defined architecture, either on the level of the binding between organic compound and cluster, or on the level of the binding between one cluster and another. This twofold lack of control nevertheless leads to objects with novel properties, such as confinement or strong fluctuations in the electromagnetic field at the optical frequencies of the plasmon resonance in the cluster. This property is commonly used in surface enhanced Raman spectroscopy (SERS) to increase the cross-section of the process, so that very low concentrations of chemical species may be detected in solution.

In contrast, when clusters are deposited from solution onto a substrate, the regular arrangement is directly induced by the surface coverage ratio and the forces governing the interaction between the clusters. For example, for gold clusters stabilised in toluene by the adsorption of bromide ions and their quaternary ammonium counterions, the simple deposition of a droplet of the solution onto a copper grid covered by a thin carbon film, followed by evaporation of the toluene, leads to a regular close-packed hexagonal arrangement of the clusters. The distance between the clusters is defined by the steric hindrance of the adsorbed organic compounds, in such a way that compounds with longer chains yield greater intercluster separations. The rigidity of the alkyl chain nevertheless remains a determining factor, as chains lose rigidity when they become too long.

Other clusters have been assembled into regular arrangements, e.g., silver or silver sulfide. The substrate plays an equally important role, since the structure of the final array results from the balance of forces between cluster and substrate, and between cluster and cluster. From the 2D array, one may



**Fig. 7.32.** Organised 2D array formed from two size distributions of gold clusters with mean diameters  $4.5 \pm 0.8$  nm and  $7.8 \pm 0.9$  nm, respectively. The gold clusters are functionalised by a thioalkane monolayer. Parameters  $p$  and  $q$  correspond to the principal crystal planes of the two size distributions [28]

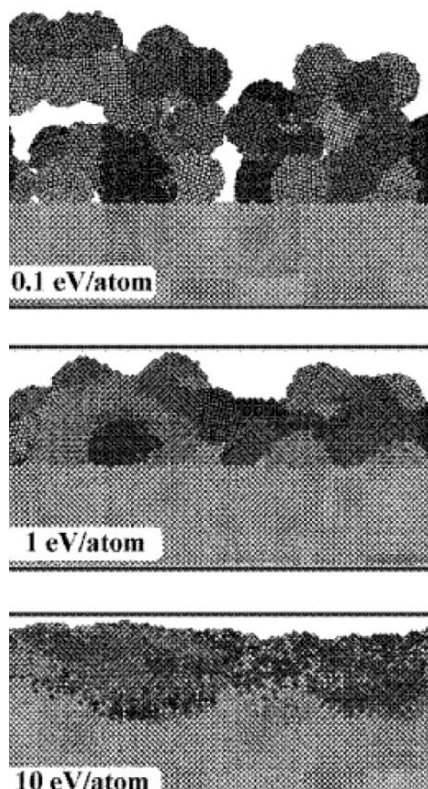
construct 3D arrays up to micrometer scales by multilayer deposition (see Fig. 7.32).

### 7.6.2 Deposition Techniques for Clusters and Colloids

It is a relatively simple matter to produce nanostructured systems by depositing small colloidal clusters previously formed in the liquid phase. The method involves depositing several droplets of the colloidal solution on a substrate and then allowing the solvent to evaporate. The clusters in suspension in the solvent will thus deposit themselves on the substrate, where they will grow into structures with different sizes and shapes depending on the nature of the substrate and the clusters, as well as the experimental conditions, especially the substrate temperature.

For systems prepared by deposition of clusters previously formed in the gas phase, the technology used is more complex. In this case, a cluster source of the kind described in Sect. 7.5.1 is placed opposite a substrate held on a mount, the whole setup being enclosed in a vacuum chamber. The cluster beam emitted by the source propagates in vacuum and the clusters are deposited on the substrate to form a thin film. Depending on the speed, and hence the kinetic energy of the clusters in the beam, which in turn depends on the type of source and the operating conditions, several characteristic deposition regimes are possible, as illustrated in Fig. 7.33.

- When the kinetic energy per atom of a cluster is less than the binding energy of the atoms making up the cluster, it will not break up on impact with the substrate. A nanoporous layer of low density will form on the substrate by random stacking of incident clusters (see Fig. 7.33). This



**Fig. 7.33.** Cross-sections obtained by molecular dynamics (MD) simulations [29] of layers of molybdenum clusters  $M_{1043}$  (diameter  $\sim 3.5$  nm), deposited at different energies on the (001) face of a molybdenum substrate. The kinetic energy of the incident clusters corresponds to 0.1 eV per atom, 1 eV per atom, and 10 eV per atom in the three simulations, viewing from *top to bottom*. For comparison, note that the binding energy between molybdenum atoms in the metal is of the order of 1.5 eV

is observed when clusters are deposited at supersonic speeds, produced naturally by an inert gas condensation source as described in Sect. 7.5.1.

- When the kinetic energy per atom of a cluster is greater than the binding energy of the atoms making up the cluster, the cluster will fragment on impact with the substrate. The fragments, which carry away part of the total kinetic energy of the incident cluster, may be implanted in the upper atomic layers of the substrate, leading to the formation of a compound. This is illustrated in the bottom picture of Fig. 7.33. Note that, from a practical point of view, high energy cluster beams are generally obtained by acceleration of ionised clusters in an electric field, such clusters being naturally produced in certain types of source. The method in which neutral clusters are ionised by electron or photon impact and then accelerated is also commonly used.
- For cluster kinetic energies between the two extreme cases mentioned above, one observes the formation of a relatively dense cluster layer, like the one shown in the central picture of Fig. 7.33.

Finally, of the three cases listed here, the one that will most concern us is the first, corresponding to a very low energy deposition regime. Indeed, when the incident clusters do not break up, we obtain a layer of nanostructured material, formed by quasi-random stacking of the basic building blocks provided by the incident clusters (see the top picture of Fig. 7.33). This material has the property of conserving the novel properties of the incident clusters. This opens the way to the synthesis of new nanostructures and nanostructured materials, with specific properties predetermined in the gas phase when the elementary building blocks, the atomic clusters, are prepared. This method, which is currently attracting a great deal of attention in the international scientific community, is referred to as low energy cluster beam deposition (LECBD).

### 7.6.3 Characteristic Mechanisms for the Formation of Nanostructures by Cluster Assembly

Since the 1980s, many experimental and theoretical studies have been carried out to understand the formation of nanostructures from clusters deposited on a substrate, in contrast to those obtained by more conventional methods based upon deposition of atoms or molecules. In the following, we shall draw heavily upon the 1999 review article [30].

To begin with, let us list the basic phenomena which may occur when clusters reach the substrate surface. At low energies (in the LE CBD regime), the clusters do not fragment upon impact with the substrate, and depending on the nature of the interaction at the cluster/substrate interface, the clusters can diffuse across the surface if the interaction is weak, or remain fixed at their point of impact if the interaction is strong. These two cases are illustrated in Figs. 7.34 and 7.35.

When gold clusters are deposited on highly oriented pyrolytic graphite (HOPG) (see Fig. 7.34), the very different natures of the two materials – the substrate is covalent and the clusters metallic – lead to a very low level of interaction between them. In fact, the  $s$  bonds in the plane of the graphite sheet and delocalisation of the  $p$  electrons from the  $p_z$  orbitals perpendicular to the plane of the sheet make this system distinctly unreactive and poorly inclined to establish bonds with atoms in the deposited gold clusters. The latter, weakly bound to the substrate, will therefore diffuse randomly over the surface until they are trapped by the various pre-existing defects on the surface, e.g., atomic steps, or associate with other clusters. This process leads to the formation of islands of clusters with a branching or ramified structure, as can be seen in Fig. 7.34.

Since the clusters are originally deposited in a quite random manner over the surface, the subsequent observation of ramified islands, in which the branches are composed by juxtaposing incident clusters, is a clear demonstration that the gold clusters diffuse over the graphite surface. The surface density of islands formed in this way can be predicted from the deposition parameters (flux of incident clusters and diffusion coefficient of the clusters



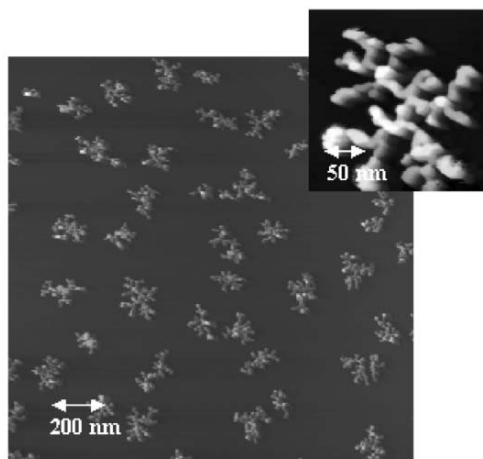
on the surface) using the DDA (deposition–diffusion–aggregation) model developed by P. Jensen [30].

In contrast, when the same gold clusters are deposited under the same conditions on a gold substrate (see Fig. 7.35), they remain isolated, distributed randomly over the surface, and this whatever the deposited thickness. In this case, the identical nature of the clusters and substrate leads to a strong interaction between them, with perfect matching of the atomic planes at the cluster/substrate interface (epitaxy). It should be remembered that the atoms on the surface of a cluster have lower coordination number than those in the bulk. They thus have a strong tendency to form bonds of the same kind as those in the bulk, in order to reduce the surface energy with respect to the bulk energy of the system. In this situation, the crystal lattice matching between cluster and substrate is a key parameter, as shown in Fig. 7.36.

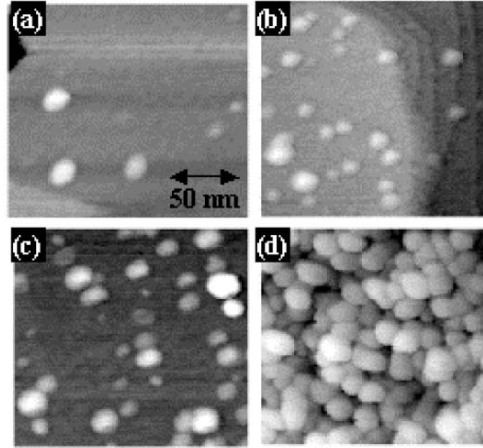
Let  $\sigma_C/\sigma_S$  denote the ratio of the crystal lattice parameters in the cluster ( $\sigma_C$ ) and the substrate ( $\sigma_S$ ). When the lattice matching between the two crystal lattices is good, we have

$$\sigma_C \approx \sigma_S . \quad (7.143)$$

This is known as the epitaxy condition. It implies an ideal situation in which cluster diffusion is completely inhibited, since the clusters remained fixed at their point of impact with the substrate. In contrast, as can be seen in



**Fig. 7.34.** Atomic force microscope image of an LECBD gold cluster deposit on an HOPG substrate at room temperature. Mean cluster size  $\text{Au}_{750}$  (diameter  $\sim 2.9$  nm). The total amount of deposited clusters corresponds to 0.3 compact monolayers of  $\text{Au}_{750}$  clusters. The easy diffusion of the gold clusters on the broad atomic terraces that are a feature of this type of graphite substrate leads to the formation of the ramified islands observed here. *Inset:* One island, showing the branches formed from strings of incident gold clusters [31]



**Fig. 7.35.** Atomic force microscope images of an LECBD gold cluster deposit (mean cluster size  $\text{Au}_{750}$  and diameter  $\sim 2.9$  nm), with different equivalent thicknesses on a gold substrate [face (111)] at room temperature. Deposited thickness (a) 0.006, (b) 0.018, (c) 0.048, and (d) 2 compact monolayers of  $\text{Au}_{750}$  gold clusters. Note that the deposited thicknesses in the first three cases are less than the thickness corresponding to the 2D percolation threshold and we observe isolated gold clusters, distributed randomly over the surface, due to their inability to diffuse in the context of the strong cluster–substrate interaction. In the last case, the deposited thickness is greater than the thickness corresponding to the 2D percolation threshold, leading to the formation of a continuous layer by random stacking of incident gold clusters [31]

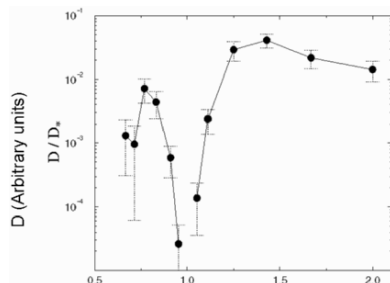
Fig. 7.36, when we move away from the epitaxy condition, cluster diffusion very quickly increases.

Regarding the way the diffusion constant  $D$  of a cluster on the substrate depends on the cluster size, measured by the number of atoms  $N$  in the cluster, the same molecular dynamics simulations mentioned for Fig. 7.36 show that  $D$  decreases when  $N$  increases, following a power law of type

$$D \sim N^{-a} . \quad (7.144)$$

The exponent  $a$ , which depends sensitively on the crystal lattice mismatch between cluster and substrate, typically varies from 0.66 for a large mismatch ( $\sigma_C/\sigma_S = 1.25\text{--}1.40$ ), to 1.4 for better matching ( $\sigma_C/\sigma_S = 1.1$ ).

The last important point we should mention here in connection with the problem of understanding and controlling the final nanostructured morphology of a cluster deposition is the coalescence of adjacent supported clusters. Indeed, depending on the extent of this phenomenon, cluster islands may evolve from branching shapes toward more compact shapes. Moreover, if we wish to conserve the memory effect with regard to the original properties of the free clusters, intercluster coalescence must be limited. It is the minimisation of the surface free energy of the system that underlies the coalescence of two



**Fig. 7.36.** Diffusion constant  $D$  of a cluster of van der Waals type (see Sect. 7.6.3) comprising 100 atoms, on a substrate of the same nature, as a function of the crystal lattice mismatch. The latter is quantified by the ratio  $\sigma_C/\sigma_S$  of the lattice parameters of the crystal lattice of the cluster ( $\sigma_C$ ) and the substrate ( $\sigma_S$ ). The graph was obtained at a temperature of  $0.3T_m$ , where  $T_m$  is the melting temperature of the material in the cluster, by a molecular dynamics simulation using an interaction potential of Lennard-Jones type [30]. The graph shows that diffusion is blocked when the epitaxy condition is satisfied for the cluster on the substrate, i.e., when  $\sigma_C/\sigma_S$  is close to unity. It also shows the high mobility of the supported cluster as soon as one moves away from this condition [32]

clusters into a single, larger and more compact cluster. In order to quantify this phenomenon, we may define a characteristic coalescence time  $t_{co}$  which will depend mainly on the nature of the clusters and the temperature.

However, it should be noted that, during the deposition of clusters on a substrate and when two adjacent clusters are attempting to coalesce, other deposited clusters diffusing on the substrate can attach themselves to this system, making it evolve towards a more ramified morphology. The characteristic time  $t_{ra}$  for this phenomenon will depend on the flux of clusters arriving at the surface, their diffusion constant, and the number of cluster islands already present.

Finally, competition will arise between the two kinetic phenomena mentioned previously, which make the system tend towards opposing morphologies – compact on the one hand and ramified on the other. The result of this competition will be determined by comparing the characteristic times of these opposing phenomena. When  $t_{ra}$  is small compared with  $t_{co}$ , ramification will win out, whereas in the other case ( $t_{ra} \gg t_{co}$ ), coalescence between clusters will be favoured, leading to more compact islands. The temperature is a parameter that may influence the result of this competition in a significant way. Indeed, when the temperature is non-negligible compared with the melting temperature  $T_m$  of the cluster material (typically  $T = 0.2T_m$ ), the kinetics of intercluster coalescence may be activated.

### 7.6.4 Examples of New Nanostructured Systems Prepared by Cluster Deposition

#### Magnetic Clusters

To our knowledge, no systems of practical interest are now produced using this method. However, the technique has some potential related to the growing need for miniaturisation of elementary components for applications in nano-electronics, nanooptics, nanomagnetism, and nanobiology, and this has led to a great deal of research in the area. These fields of applications are described in detail in Chap. 4. However, this is a good point to discuss some of the specific prospects offered by cluster techniques in the design of such futuristic systems.

In particular, one of the essential requirements in the context of these applications concerns the preparation of a new generation of systems with very high integration density (typically  $10^{11}$ – $10^{12}$  nano-objects per  $\text{cm}^2$ ). A standard example in this area is the problem of high density magnetic recording, such as the hard disk of a computer. Recording densities have currently reached ten to a few tens of  $\text{Gbits}/\text{cm}^2$  and are soon expected to move into the range of  $100 \text{ Gbits}/\text{cm}^2$  to  $1 \text{ Tbits}/\text{cm}^2$ . To achieve this, one must overcome the physical limit related to superparamagnetic behaviour at room temperature which is a characteristic of nanoscale magnetic systems (see Chap. 4). This behaviour occurs when

$$KV < k_{\text{B}}T . \quad (7.145)$$

Indeed, recall that the anisotropy energy of a magnetic cluster (proportional to  $KV$ , where  $K$  is the anisotropy constant of the magnetic material and  $V$  the cluster volume) is what causes the magnetisation of the cluster to block in one of the easy magnetisation directions. However, when  $V$  decreases, this anisotropy energy can become small enough to suffer competition from the thermal energy  $k_{\text{B}}T$ . The latter tends to shift the magnetisation of the cluster between the easy directions. A cluster of this kind cannot be used as an elementary magnetic recording dot, because such an entity requires the magnetisation to remain blocked in the writing direction.

One solution would be to increase the magnetic anisotropy  $K$  of the cluster by a corresponding amount, to counterbalance its decreased volume  $V$ , in such a way that  $KV$  always remains above the thermal energy  $k_{\text{B}}T$  ( $\sim 25 \text{ meV}$  at room temperature). Among the various sources of magnetic anisotropy in a cluster, one may try to increase the magnetocrystalline component of the anisotropy by modifying the material making up the cluster. However, it is no easy matter to fabricate stable magnetic nano-alloys with well controlled composition using conventional methods, whereas such materials may be more accessible when clusters are formed in the gas phase, in non-equilibrium conditions, for example.

Hence, for applications to magnetic nanostructures, magnetic clusters or colloids made from transition metals such as Co, Fe or Ni were synthesised to begin with and subjected to wide-ranging investigation. Since the magnetic blocking temperatures determined by the magnetic anisotropy of the materials in these nanosystems are too low for practical applications (from ten to a few tens of kelvins for clusters containing a few hundred to a few thousand atoms, i.e., diameters 2–10 nm), research soon turned to bimetallic clusters and colloids, e.g., Co-Pt, Co-Sm, Fe-Pt, which have much higher magnetic anisotropies, leading to magnetic blocking temperatures higher than room temperature in some cases. It should be noted that several mechanisms lead to enhancement of the magnetic anisotropy in small bimetallic clusters and they are not always very well understood. Both bulk and surface effects may be involved. In the latter case, the segregation phenomenon whereby the element which confers the lowest surface energy on the cluster surface tends to concentrate there, may lead to clusters with a heterogeneous morphology, often referred to as core-shell clusters.

### Covalent Clusters

Besides the characteristic example of magnetic nanosystems described above, which represent a major trend in technology for the present decade, one should also mention the example of semiconductor nanostructures, which are destined to generate key technological developments in the move from the microelectronics of today to the nanoelectronics of tomorrow. In this area, novel and interesting prospects are offered by deposition of clusters that have been preformed in the gas phase, and these methods are currently under investigation. In particular, certain covalent semiconductor materials already well known in the world of microelectronics, such as silicon, can exhibit completely new structures and electronic properties on the nanoscale.

This is especially true of the cage-type clusters described in Chap. 8. One well known example, made from carbon, is provided by the fullerenes, the best known being  $C_{60}$  with its familiar football appearance. In the case of silicon, fullerene-type clusters (hollow cages) are observable at small sizes (up to  $Si_{28}$ ), whereas stuffed fullerenes tend to form at larger sizes to allow the silicon atoms to conserve their  $sp^3$  hybridisation. However, in all cases, the highly specific atomic structure of these cage clusters compared with the diamond phase of bulk solid silicon, which contains among other things a large number of pentagonal cycles, confers electronic structures and resulting properties upon these semiconductor nanosystems that are quite unique. In particular, an almost direct electronic band gap, much broader than that of bulk silicon ( $\sim 1.2$  eV), causes a photoluminescence effect in the visible. This should allow silicon to make its first appearance in the field of optoelectronics, whilst bulk solid silicon with its narrow and indirect band gap is unsuited to this type of application, generally fulfilled by semiconductors of type III-V (GaAs, InP).

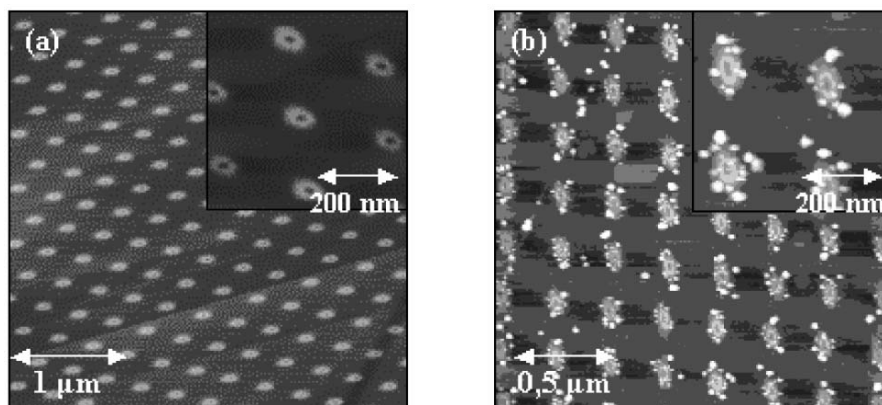
### Cluster Arrays

In the two examples of magnetic or semiconductor nanostructures discussed above, the preformed cluster channel is the most widely used to synthesise functional nano-objects, with a high magnetic anisotropy in one case, or a photoluminescence property in the other, in order to satisfy the requirements of certain applications. However, if one is to build any kind of component, then following this first stage of functionalisation of the nanosystem, one must assemble the basic nanosystems in an ordered manner to implement addressing. This can be envisaged by controlling the nucleation and growth of nano-objects deposited on a substrate, itself suitably functionalised in a previous step. For clusters that have been preformed in the gas phase and deposited on a substrate in the LECBD regime as described in Sect. 7.6.2, one simply uses a substrate patterned with a regular array of traps for the diffusing clusters.

As an example, consider the case discussed in Sect. 7.6.3 and Fig. 7.34 in which gold clusters  $\text{Au}_{750}$  were deposited on an HOPG substrate. In this case, the random deposition and diffusion of the clusters led to the formation of randomly distributed islands on the surface. Using nanoetching techniques (see Chap. 1), an ordered array of artificial defects can be inscribed on the surface of the graphite substrate, as shown in Fig. 7.37a. Gold clusters deposited on this substrate will be trapped by these artificial defects as they diffuse over the surface. At the end of the deposition, an ordered 2D array of gold cluster islands will be produced, as shown in Fig. 7.37b. Ordered nanoparticle arrays can also be fabricated from colloidal solutions (see Sect. 7.6.1 and Fig. 7.30).

Finally, there are many other applications using clusters or colloidal particles. For example, in the field of chemical catalysis, finely divided materials are needed, whilst in the study of biological vectors, clusters could be used as vehicles for transporting various surface-grafted active molecules towards some predetermined target in a living organism. It would be difficult to catalogue all these applications here and it seems preferable to try to identify the general nature of the nanostructured materials prepared by assembling clusters or colloids.

This general character depends for the main part on the combination of or competition between the inherent properties of the clusters and the interactions between neighbouring clusters. This has been clearly demonstrated in the case of nanostructured layers of magnetic cobalt, iron or nickel clusters, for which a behaviour of correlated spin glass type is observed, intermediate between the behaviour of amorphous magnetic and classic ferromagnetic materials (see Chap. 4). It should be noted that the intrinsic properties of clusters are conditioned at the moment they are prepared in the gas or liquid phase, whereas the interactions between clusters will depend on the deposition conditions and the conditions of nucleation and growth on the substrate, which can influence the final nanostructured morphology of the system. These two stages, i.e., the initial synthesis of the nano-objects, followed by deposition on a substrate, are completely independent and involve completely unrelated

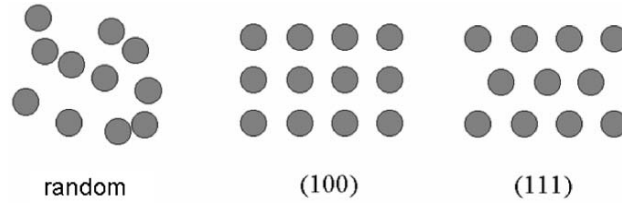


**Fig. 7.37.** (a) Atomic force microscope image of an ordered array of nanoscale defects inscribed on an HOPG substrate by ion beam bombardment. A focussed beam of  $\text{Ga}^+$  ions of energy 30 keV is used in this case, with a dose of 37 500  $\text{Ga}^+$ /point, producing defects with nanocrater-type morphology. This is clearly visible in the highly magnified image shown in the *insert*. The period of the array, i.e., the distance between two consecutive nanocraters, is 300 nm. Note that the morphology of the defects can be rather tightly controlled through the energy and type of ions used, as well as the ion dose per point. (b) Atomic force microscope image of an LECBD deposition of gold clusters with mean size  $\text{Au}_{750}$  and diameter  $\sim 2.9$  nm on a nanoetched HOPG substrate of the type shown in (a). The amount of gold clusters deposited corresponds to  $10^{-2}$  compact monolayers. The substrate temperature during deposition is held at  $100^\circ$  to facilitate diffusion of the deposited clusters. The magnified image in the *insert* clearly shows the gold clusters captured at the edges of the nanocrater-type defects. This example shows that the nucleation and growth of cluster islands can be controlled on a functionalised substrate in order to fabricate components with very high surface integration densities, typically in the range  $10^2$ – $10^3$  Gbits/ $\text{cm}^2$  [33]

control parameters. This introduces a high level of versatility in the choice of parameters and experimental conditions, making these techniques ideal for the preparation of novel nanostructures that could not be achieved by more conventional techniques based on atomic or molecular deposition, for example, since these do not allow one to disconnect the different stages of the synthesising process.

### Collective Properties

To a large extent, current interest in cluster assemblies rests upon the fact that they have different properties to the isolated clusters. Hence, when metallic clusters are deposited on an insulating substrate by the methods discussed above, one observes a transition from an insulating film to a conducting film



**Fig. 7.38.** 2D array of metallic clusters with variable topology: random, (100) and (111)

by modifying the surface cluster density: this is the Mott insulator–metal transition.

The electron density of a spherical metallic cluster extends beyond the size of the cluster as defined by its ionic core by a distance of the order of the Fermi wavelength  $\lambda_F$ . In a sufficiently dense assembly of clusters, electron spillover from neighbouring clusters can overlap, whereupon the electron density is delocalised over the whole assembly. This regime is only achieved for small intercluster separations, when the clusters are close to contact. For larger intercluster separations, the film is insulating, because the clusters are then isolated, without coupling to their nearest neighbours, so that the electron density is localised on the clusters. The electrical properties of these films can thus be controlled simply by controlling the surface density of the clusters. Current work aims to investigate the exact role of the organisation of these clusters in specific patterns of the type observed in the (100) or (111) crystallographic planes of a cubic system, for example, as shown in Fig. 7.38.

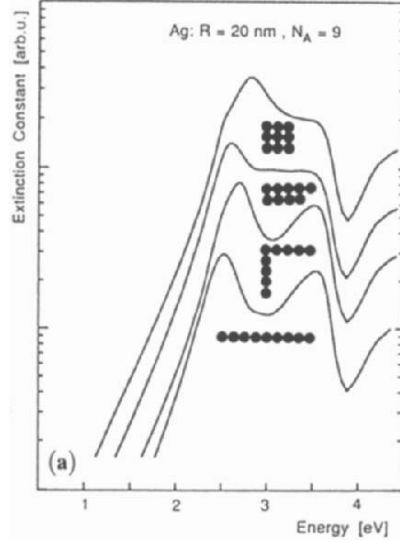
Concerning optical properties, the changes to the absorption spectrum, for example, are very significant due to the delocalisation of conduction electrons over the whole ensemble of clusters, rather than just a single cluster. In particular, new resonances are observed in the absorption spectrum at visible frequencies for gold or silver clusters, at energies defined by the topology of the cluster assembly (see Fig. 7.39).

From a formal standpoint, the dielectric constant at optical frequencies of an assembly of clusters with dielectric constant  $\varepsilon$  embedded in a matrix with dielectric constant  $\varepsilon_m$  can be approximated by the Maxwell–Garnett or Bruggeman model:

$$f \frac{\varepsilon - \varepsilon_f}{\varepsilon + 2\varepsilon_f} + (1 - f) \frac{\varepsilon_m - \varepsilon_f}{\varepsilon_m + 2\varepsilon_f} = 0, \quad (7.146)$$

where  $f$  is the volume fraction of metal and  $\varepsilon_f$  is the dielectric constant of the film. This expression cannot correctly account for the nanostructuring of the film, or for its exact topology, the latter only being introduced in the form of a homogeneous statistical distribution. When the separation between clusters is too small, this model deviates significantly from experimental results obtained by observing the absorption spectrum of the film. This expression can thus





**Fig. 7.39.** Extinction spectrum of nine silver clusters in assemblies with different topologies, calculated using the generalised Mie theory [15]

only be used for volume fractions well below 0.01, although a discrepancy is already observed for values below this limit.

The full theory of the optical response of assemblies containing a large number of clusters is the generalised Mie theory, which specifically takes into account the topological arrangement of the clusters in order to account for multiple diffusion. A less sophisticated approach consists in restricting to the electric dipole approximation, and hence to clusters with diameters well below the incident wavelength. Each cluster  $i$  is replaced by the corresponding induced dipole, treated as a point. The coupling between the different dipoles is then modelled by an induced dipole–induced dipole interaction term. The  $\alpha$  component of the dipole  $p_i$  of the  $i$ th cluster is then given by

$$p_{i\alpha} = \alpha_0 \left[ E_{\alpha}^{(0)} + \sum_{j=1, j \neq i}^N \sum_{\beta} G_{ij, \alpha\beta} p_{j\beta} \right], \quad (7.147)$$

where  $\alpha_0$  is the polarisability of the cluster and  $\mathbf{E}^{(0)}$  is the electric field of the incident wave. The second term in the square brackets is the induced dipole–induced dipole coupling, given by

$$G_{ij, \alpha\beta} = \frac{3r_{ij, \alpha} r_{ij, \beta} - \delta_{\alpha\beta} r_{ij}^2}{r_{ij}^5} \quad (7.148)$$

in the quasi-static approximation, i.e., by neglecting retardation effects due to the finite speed of the electromagnetic wave. In these equations, the vectorial

nature of the problem has been shown by explicitly introducing the different components  $\alpha = x, y, z$  of the dipoles and the distance  $r_{ij}$  between the  $i$ th and  $j$ th clusters. The symbol  $\delta_{\alpha\beta}$  is the Kronecker delta. One thus obtains a system of  $3N$  coupled equations which can be solved quickly if the number of clusters  $N$  is not too high.

## 7.7 Conclusion and Prospects

The examples given above to illustrate the size effects and intrinsic properties arising from a finite number of atoms should convince the reader of the great potential of nanostructured systems. Nanotechnology requires specific tools for fabricating materials, such as LECBD deposition, laser vaporisation sources, colloidal systems, etc., and tools for characterising and manipulating them, such as scanning tunneling microscopy, combined with a reasoned approach to the properties of the elementary building blocks themselves. Indeed, it would be a mistake to consider only the scale factor, when the properties of these elements are so different from our intuitive ‘school book’ understanding.

The reader must therefore develop a familiarity with novel ideas: a metal becoming an insulator, or an inert material becoming explosive due to an ultrahigh level of reactivity, for example. Nothing is obvious, everything is possible: this is the richness of nanostructured systems. The reasoned approach consists therefore in studying the influence of the finite number of atoms through the surface concept ( $1/R$  law) and the so-called molecular approach illustrated in this chapter. The discovery of the memory effect in films, wherein a thin film formed by depositing elementary building blocks conserves to some extent the properties of those building blocks, opens up a whole new area of physics in which properties can be adapted at will by playing with the size of the basic elements.

### Acknowledgements

The authors would like to thank M. Boyer and P. Jensen for many fruitful discussions which have helped in the preparation of this chapter.

## Appendix

### A. Polyhedra

The cube, octahedron, cubo-octahedron with hexagonal faces and cubo-octahedron with triangular faces can be generated by changing the ratio of the distances of the (111) and (100) facets from the centre:

Cube	$\frac{R_{111}}{R_{100}} = \sqrt{3}$
Cubo-octahedron with hexagonal faces	$\frac{R_{111}}{R_{100}} = \frac{\sqrt{3}}{2}$
Cubo-octahedron with triangular faces	$\frac{R_{111}}{R_{100}} = \frac{2}{\sqrt{3}}$
Octahedron	$\frac{R_{111}}{R_{100}} = \frac{1}{\sqrt{3}}$

It is clear that the cube and octahedron have a high degree of anisotropy according to the Wulff criterion.

### B. Hybridisation

Hybridisation occurs by linear combination of pure  $s$  and  $p$  orbitals. The tetravalent elements (one  $s$  electron and three  $p$  electrons) offer three possibilities:

- hybridisation  $sp_x p_y, p_z$  dihedral angle  $180^\circ$  (linear structure),
- hybridisation  $sp_{x,y}^2 p_z$  dihedral angle  $120^\circ$  (hexagonal structure),
- hybridisation  $sp_{x,y,z}^3$ , dihedral angle  $109^\circ 47'$  (tetrahedral structure).

### C. Cohesive Energy

In the liquid-drop model (7.8):

$$E_{\text{cluster}}/N = E_{\text{cohesive}}(1 - 0.82N^{-1/3}) . \quad (7.149)$$

In the second moment approximation (7.16):

$$E_{\text{cohesive}} = a \times C_{\text{max}}^{1/2} , \quad (7.150)$$

where  $a$  is a coefficient of proportionality. Hence,

$$E_{\text{cluster}} = aC^{1/2} , \quad (7.151)$$

and consequently,

$$C = C_{\text{max}}(1 - 0.82N^{-1/3})^2 . \quad (7.152)$$

### D. Work Function

The exact calculation is tedious and difficult. The details can be found in [34, 35]. Suppose that an electron is torn from an infinite metal surface ( $R \rightarrow \infty$ ). At a distance  $x$  from the surface, this electron is subject to a Coulomb interaction between itself and the hole of charge  $+q$  delocalised on the surface (hence equivalent to a positively charged surface). It can be shown that the

interaction energy can be written using the image force model. In this model, there is a charge  $-q$  (the electron) at a distance  $d$  from the surface and a fictitious charge  $+q$  at a distance  $-d$  within the surface, so that

$$U_{\infty} = \frac{q^2}{16\pi\epsilon_0 x}, \quad (7.153)$$

where  $\epsilon_0$  is the vacuum permittivity and  $x$  is the distance from the electron to the surface. In the case where  $x = 1 \text{ \AA}$ , a characteristic interatomic distance,  $U_{\infty} = 3.6 \text{ eV}$  is the order of magnitude of the energy required to remove an electron from a previously neutral conducting surface. This is nothing other than the work function  $\Phi$ . Note also that  $U_{\infty}$  is half the interaction energy between two real charges located at a distance  $2x$  apart. In our case, we have only one charge.

For a sphere of finite radius  $R$ , it can be shown that the potential is

$$U_R = \frac{q^2 R^3}{8\pi\epsilon_0 (R+x)^2 [(R+x)^2 - R^2]} + \frac{q^2}{4\pi\epsilon_0 (R+x)}. \quad (7.154)$$

If we calculate the difference between the potential of a surface [sphere of infinite radius (7.153)] and the potential of a droplet of radius  $R$  given by (7.154), then setting  $x = R\delta$ , we have

$$\Delta U = U_{\infty} - U_R = \frac{q^2}{4\pi\epsilon_0 R(1+\delta)} + \frac{q^2}{8\pi\epsilon_0 R\delta(1+\delta)^2(2+\delta)} - \frac{q^2}{16\pi\epsilon_0 R\delta}. \quad (7.155)$$

There is no condition on  $x$ . We may choose a value as small as possible, such that  $x \ll R$ , i.e.,  $\delta \rightarrow 0$ . It then follows that, to first order in  $\delta$ ,

$$\Delta U = \frac{3q^2}{32\pi\epsilon_0 R}. \quad (7.156)$$

We may then define the ionisation potential of the droplet by

$$PI(R) = \Delta U + \Phi, \quad (7.157)$$

or

$$PI(R) = \Phi + \frac{3q^2}{32\pi\epsilon_0 R}, \quad (7.158)$$

or again,

$$PI(R) = \Phi + \frac{5.4}{R} \text{ eV}, \quad (7.159)$$

where  $R$  is expressed in angstrom units.

## E. Kubo Criterion

### *Transition Metal Clusters ( $d$ Electrons)*

The same argument can be used with the transition metals, taking into account the specific form of the  $d$  orbitals, which are much more localised in space (cigar shape) than the  $s$  electrons (sphere).

The transition metals are characterised by the presence of a  $d$  band much narrower than the  $s$  band. This arises because of the much stronger localisation of the  $d$  orbitals. As a consequence, the density of states at the Fermi level will be higher, and the Kubo transition will occur at much smaller sizes. There is no closed analytic expression to describe the transition, because the globally rectangular  $d$  band may manifest a certain number of anomalies depending on the crystallographic structure:

- For the centered cubic structure, which has two maxima and a reduction in the density of states at the band center. This effect is analogous to the forbidden band observed in semiconductors, although the two bands are occupied.
- Fine structures made up of narrow but intense peaks in the vicinity of  $E_F$  in the face-centered cubic structure.

### *Clusters of Divalent Metals ( $s$ and $p$ Electrons)*

The divalent elements Hg, Zn, etc., have a closed  $s$  shell. According to the free electron model, such elements should lead to completely filled bands and hence to insulating structures. This is not the case, because the divalent elements are metallic in the bulk solid phase. The electron delocalisation arises from the increasing overlap of the broadening  $s$  and  $p$  bands as  $N$  increases, e.g., in mercury. The width of an  $s$  or  $p$  band is given in the second moment approximation of the tight-binding model (TBSMA) by

$$W_{s,p} = a_{s,p} C^{1/2}, \quad (7.160)$$

where  $C$  is the number of nearest neighbours and  $a$  a constant function of the type of orbital. As the  $p$  orbitals are more localised than the  $s$  orbitals, the band width  $W$  will be narrower for the  $p$  states ( $a_p < a_s$ ). If we do not take into account the hybridisation between the  $s$  and  $p$  states, which leads to a covalent character, rather than a metallic one, the insulator–metal transition will occur when (see Fig. 7.8)

$$a(W_s + W_p) \approx E_p - E_s, \quad (7.161)$$

where  $E_s$  and  $E_p$  are the positions of the atomic  $s$  and  $p$  levels and  $a$  is a constant between 1/2 and 1.

We noted in Appendix C that the mean coordination number is related to the cluster size by [see (7.152)]

$$C = C_{\max}(1 - 0.82N^{-1/3})^2, \quad (7.162)$$

where  $C_{\max}$  is the maximal coordination number in the solid phase. Taking the example of mercury, the width of the  $sp$  band in the solid is essentially fixed by the  $s$  band, which implies that

$$W_{s,p}(\text{solid}) = a_{s,p}C_{\max}^{1/2} = 9.2 \text{ eV}. \quad (7.163)$$

Hence, for a cluster of  $N$  atoms, we have a band width

$$W_{s,p} = 9.2(1 - 0.82N^{-1/3}). \quad (7.164)$$

$E_s$  and  $E_p$  are the positions of the atomic levels  $6s$  and  $6p$  with  $E_p - E_s \sim 5 \text{ eV}$ . According to (7.161) and (7.164), one thus expects a transition for sizes  $N_{\text{crit}}$  given by

$$W_{s,p} = 9.2(1 - 0.82N_{\text{crit}}^{-1/3}) \sim 10a. \quad (7.165)$$

This model is not accurate enough to calculate  $N_{\text{crit}}$ . Experimentally, it is found that the transition occurs above 13 atoms ( $a = 0.6$ ). It should be noted that the band overlap does not depend on  $k_B T$ , and the Kubo criterion is not the right parameter for this type of transition.

#### *Clusters of Covalent Elements ( $s$ and $p$ Electrons)*

These elements have partially occupied  $p$  states, where the configuration of the valence electrons is  $s^2p^2$ . We shall only consider 3D structures. In this case, bonding and anti-bonding states created by hybridisation of  $s$  and  $p$  states are separated by a forbidden band or gap of width  $E_g$ . This hybridisation is called  $sp^3$  hybridisation. In the bulk solid phase, the excitation of an electron from the valence band to the conduction band generates a hole in the valence band. The electron-hole interaction is called an exciton. It can be shown that the energy of this exciton in the Bohr model is given approximately by

$$E_{\text{exc}} = -\frac{q}{8\pi\epsilon_0\epsilon a_B}, \quad (7.166)$$

where  $a_B$  is the Bohr radius ( $0.527 \text{ \AA}$ ) and  $\epsilon$  the static dielectric constant of the material. The radius  $a_B^*$  of the associated Bohr orbit is about  $4.3 \text{ nm}$  for silicon ( $\epsilon = 12$ ). If the cluster is smaller in size than the Bohr orbit, the exciton will be quantised. One may then apply the model of a particle confined within a box. For a square potential well, we have the well known formula

$$E_n = \frac{1}{2m} \left( \frac{n\hbar\pi}{L} \right)^2, \quad (7.167)$$

where  $L$  is the length of the box. If  $L \ll a_B^*$ , the electron and hole will be quantised separately, which leads to

$$E_n = \frac{1}{2m_e^*} \left( \frac{n\hbar\pi}{L} \right)^2 + \frac{1}{2m_h^*} \left( \frac{n\hbar\pi}{L} \right)^2, \quad (7.168)$$

where  $m_e^*$  and  $m_h^*$  are the effective masses of the electron and hole, respectively. In this case, the energy that must be supplied to the electron will be given by

$$E_{\text{tot}} = E_g + E_n. \quad (7.169)$$

This is the well known confinement model. When  $L$  decreases (the cluster volume is  $L^3$  here),  $E_{\text{tot}}$  increases, indicating that the levels move apart from one another. This is a direct consequence of the Kubo criterion. Taking  $n = 1$ ,  $m_e^* = m_h^* = m$ , it then follows that

$$E_{\text{tot}} \text{ (eV)} = E_g + \frac{0.75}{L^2 \text{ (nm)}}. \quad (7.170)$$

The model fails when the cluster size becomes too small. For very small sizes, the  $sp^3$  hybridisation is no longer the stable phase due to the presence of a great many dangling bonds at the cluster surface. If atomic states are accessible, e.g.,  $d$  states, hybridisation will occur between the  $s$ ,  $p$  and  $d$  states, with an increase in the mean coordination ( $dsp^3$  hybridisation forms a tetragonal bipyramid). In this case, the forbidden band disappears and there is a transition towards a metallic state. This phase transition is predicted for elements in column IVA, with the exception of carbon, where the  $d$  states are inaccessible.

## F. Magnetic Susceptibility

We calculate here the logarithm and successive derivatives. Hence,

$$\ln(Z_{\text{even}}) = -\frac{E_0}{k_B T} + \ln \left[ 1 + 2 \exp \left( -\frac{\Delta}{k_B T} \right) \left( 1 + \cosh \frac{2\mu_B H}{k_B T} \right) + \exp \left( -\frac{2\Delta}{k_B T} \right) \right], \quad (7.171)$$

$$\frac{\delta \ln(Z_{\text{even}})}{\delta H} = \frac{\frac{4\mu_B}{k_B T} \exp \left( -\frac{\Delta}{k_B T} \right) \sinh \frac{2\mu_B H}{k_B T}}{1 + 2 \exp \left( -\frac{\Delta}{k_B T} \right) \left( 1 + \cosh \frac{2\mu_B H}{k_B T} \right) + \exp \left( -\frac{2\Delta}{k_B T} \right)}, \quad (7.172)$$

and

$$\frac{\delta^2 \ln(Z_{\text{even}})}{\delta H^2} = \frac{8 \left( \frac{\mu_B}{k_B T} \right)^2 \exp \left( -\frac{\Delta}{k_B T} \right) \cosh \frac{2\mu_B H}{k_B T}}{1 + 2 \exp \left( -\frac{\Delta}{k_B T} \right) \left( 1 + \cosh \frac{2\mu_B H}{k_B T} \right) + \exp \left( -\frac{2\Delta}{k_B T} \right)}. \quad (7.173)$$

Then as  $H \rightarrow 0$ ,

$$\left. \frac{\delta^2 \ln(Z_{\text{even}})}{\delta H^2} \right|_{H \rightarrow 0} = 8 \left( \frac{\mu_B}{k_B T} \right)^2 \frac{\exp(-\Delta/k_B T)}{1 + 4 \exp(-\Delta/k_B T) + \exp(-2\Delta/k_B T)}. \quad (7.174)$$

The susceptibility is obtained from (7.66) and (7.174) by integrating over all possible states  $\Delta$ , whence

$$\chi_{\text{even}} = \frac{8\mu_B^2}{k_B T} \int \frac{P_0(\delta) \exp(-\Delta/k_B T)}{1 + 4 \exp(-\Delta/k_B T) + \exp(-2\Delta/k_B T)} d\Delta. \quad (7.175)$$

Using (7.63), it follows that

$$\chi_{\text{even}} = \frac{8\mu_B^2}{k_B T} \int \frac{\exp(-\Delta/\delta) \exp(-\Delta/k_B T)}{\delta [1 + 4 \exp(-\Delta/k_B T) + \exp(-2\Delta/k_B T)]} d\Delta. \quad (7.176)$$

This integral has been tabulated numerically by Denton et al. and we obtain

$$\chi_{\text{even}} = 3.04\mu_B^2/\delta = N(E_F)3.04\mu_B^2. \quad (7.177)$$

Now consider

$$\frac{\delta \ln(Z_{\text{odd}})}{\delta H} = (\mu_B/k_B T) \frac{\sinh(\mu_B/k_B T)}{\cosh(\mu_B/k_B T)}, \quad (7.178)$$

and hence

$$\frac{\delta^2 \ln(Z_{\text{odd}})}{\delta H^2} = \frac{(\mu_B/k_B T)^2}{\cosh^2(\mu_B/k_B T)}. \quad (7.179)$$

Then for  $H \rightarrow 0$ ,

$$\left. \frac{\delta^2 \ln(Z_{\text{odd}})}{\delta H^2} \right|_{H \rightarrow 0} = \left( \frac{\mu_B}{k_B T} \right)^2. \quad (7.180)$$

The susceptibility is obtained from (7.66) and (7.180) by integrating over all possible states  $\Delta$ , whence

$$\chi_{\text{odd}} = \frac{\mu_B}{k_B T} \int P_n(\Delta) d\Delta. \quad (7.181)$$

Integrating the energy distribution  $P_n(\Delta)$  over all possible  $\Delta_n$  gives unity. The susceptibility  $\chi_{\text{odd}}$  is then

$$\chi_{\text{odd}} = \mu_B^2/k_B T. \quad (7.182)$$



### G. Fermions and Magic Sizes

Fermions are elementary particles with half-integer spin, e.g., electrons, protons, neutrons, the particles making up ordinary matter. Fermions obey Fermi–Dirac statistics and the Pauli exclusion principle, which says that two identical fermions, such as two electrons, cannot occupy the same quantum state. Most properties of matter are consequences of this principle, including atomic structure, molecular structure, electronic band structure in solids, and so on.

Magic numbers (or sizes) are specific numbers of fermions (electrons in the case of an atom or cluster, nucleons in the nucleus) at which the system has greater stability than at adjacent sizes. The terminology comes from nuclear physics, where certain nuclei with specific numbers of protons and neutrons are extremely stable.

### H. Time-of-Flight Spectrometer

The clusters emerging from a neutral cluster source generally have almost the same velocity. To find the distribution of sizes  $N$  (or masses  $M$ ) in the cluster beam, they are ionised (photoionisation by laser, or bombardment by an electron gun) as they enter a device known as the time-of-flight spectrometer. This comprises a small acceleration region in which there is an intense static electric field, followed by a long region (length  $L$ ) called the free flight zone. In the first part, the ionised clusters are accelerated by the intense electric field and reach different speeds  $v$  depending on their mass  $M$  ( $v \propto M^{-1/2}$ ). In the free flight zone, the clusters propagate at constant speed, i.e., the speed acquired at the end of the acceleration region, and their time of arrival  $t$  at a detector placed at the end of the free flight zone will therefore be a function of their mass ( $t \propto LM^{1/2}$ ). A specific mass thus corresponds to each time of arrival, and the intensity of the detected signal will reflect the number of clusters having this mass (hence the generic name of mass spectrometer).

### I. Sequence of Magic Sizes

The main electronic magic sizes  $N_e$  (20, 40, 58, 92, 138, etc.) are not regularly spaced. However, their cube roots are. This result is intuitively obvious as regards the atomic shell structure, since the transition from one perfect polyhedron to the next corresponds to a covering of the surface (or the half surface, as in the case of an octahedral geometry) by a further ‘layer’ of atoms. It indicates that the transition from one electronic magic size to the next corresponds to a fixed increase in the cluster radius. This surprising characteristic is explained by the semiclassical theory of the density of states, discussed in Sect. 7.4.2.

### J. Negligible Trajectories

The almost non-existent influence of the oscillating trajectory  $M(1, 2)$  (in fact, the shortest) arises because the quantum states  $E_{n,l}$  contributing most to a component  $D_N^M(E_{\text{kin}}^{1/2})$  are those whose angular momentum  $[l(l+1)]^{1/2}\hbar$  is of the order of magnitude of the angular momentum  $L$  characterising the classical orbit. [The amplitude factor  $A_M$  is in fact proportional to  $(L + \hbar/2)$ .] The component  $D_N^{\text{fluc}}(E_{\text{kin}}^{1/2})$  is thus essentially dominated by the short orbits with high angular momentum. The ‘polygonal’ trajectories making one round trip  $M(q=1, n>4)$  (pentagon, hexagon, etc.) also have little relative influence because, for one thing, they do not exist over a large size range due to edge effects [in the presence of an effective potential  $V_{\text{eff}}(r)$  with edges that are not too abrupt, the rebound against the surface requires a minimal angular rotation], and for another thing, their lengths (parameters  $\alpha_M$ ) are relatively close to the length of the square orbit. Calculation shows that these trajectories can be taken into account by attributing a very low apparent anharmonicity to the contribution of the square orbit.

### K. Basic Electromagnetism of Homogeneous, Isotropic Media in the Linear Approximation

#### Static Regime

In the presence of a macroscopic electric field  $\mathbf{E}(\mathbf{r})$ , the atoms or molecules in the medium are polarised. In fact, each atom or molecule can be treated as a small electric dipole moment, or a point dipole. The electric dipole density at  $\mathbf{r}$ , i.e., the sum of all the small dipoles in the element of volume  $d^3\mathbf{r}$  centered on  $\mathbf{r}$ , divided by  $d^3\mathbf{r}$ , is the polarisation vector  $\mathbf{P}(\mathbf{r})$  given by

$$\mathbf{P}(\mathbf{r}) = \varepsilon_0\chi\mathbf{E}(\mathbf{r}), \quad (7.183)$$

where  $\chi$  is the linear electric susceptibility of the medium. The electric displacement vector  $\mathbf{D}(\mathbf{r})$  is defined by

$$\mathbf{D}(\mathbf{r}) = \varepsilon_0\mathbf{E}(\mathbf{r}) + \mathbf{P}(\mathbf{r}) = \varepsilon_0(1 + \chi)\mathbf{E}(\mathbf{r}) = \varepsilon_0\varepsilon\mathbf{E}(\mathbf{r}), \quad (7.184)$$

where  $\varepsilon$  is the relative dielectric constant (or permittivity) of the medium. If there are no free charges,  $\text{div}(\mathbf{D}) = 0$ . Since the electric field and electrostatic potential  $V(\mathbf{r})$  are related by

$$\mathbf{E}(\mathbf{r}) = -\text{grad}[V(\mathbf{r})], \quad (7.185)$$

we have the Poisson equation

$$\Delta V(\mathbf{r}) = 0, \quad (7.186)$$

where  $\Delta$  is the Laplacian.

### Time-Varying Regime

When there is a monochromatic oscillating electric field  $\mathbf{E}(\mathbf{r}) \cos(\omega t)$ , the polarisation vector  $\mathbf{P}(\mathbf{r}, t)$  oscillates with angular frequency  $\omega$  in the direction of the field. However, there is generally a phase difference with respect to the field, i.e., it has a component in phase quadrature with the field. It is this component which is responsible for energy absorption by the medium, e.g., the incident electromagnetic energy is converted into heat. The amplitudes of the components in phase and in phase quadrature depend on the angular frequency  $\omega$ . As in electrokinetics, it is better to use a complex representation in order to avoid tedious trigonometric calculations. The complex representations given by

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}(\mathbf{r}) \exp(-i\omega t), \quad \mathbf{P}(\mathbf{r}, t) = \mathbf{P}(\mathbf{r}) \exp(-i\omega t),$$

$$\mathbf{D}(\mathbf{r}, t) = \mathbf{D}(\mathbf{r}) \exp(-i\omega t), \quad V(\mathbf{r}, t) = V(\mathbf{r}) \exp(-i\omega t),$$

are therefore used in all algebraic manipulations. The true physical quantities are, of course, the real parts of these quantities.  $\mathbf{E}(\mathbf{r})$ ,  $\mathbf{P}(\mathbf{r})$ ,  $\mathbf{D}(\mathbf{r})$  and  $V(\mathbf{r})$  are thus complex quantities. The following complex quantities are then defined in the complex representation:

$$\chi(\omega) = \chi_1(\omega) + i\chi_2(\omega), \quad (7.187)$$

$$\varepsilon(\omega) = \varepsilon_1(\omega) + i\varepsilon_2(\omega), \quad (7.188)$$

where  $\varepsilon(\omega)$  is the complex dielectric function of the medium.

We have  $\chi_2(\omega) = \varepsilon_2(\omega) = 0$  when the medium is non-absorbent (zero component in phase quadrature). The polarisation current density  $\mathbf{j}$  is related to  $\mathbf{P}$  by

$$\mathbf{j}(\mathbf{r}, t) = \frac{\partial \mathbf{P}(\mathbf{r}, t)}{\partial t}. \quad (7.189)$$

The power delivered to the medium by the field (heat production) in a volume element  $d^3\mathbf{r}$  is  $\mathbf{j} \cdot \mathbf{E} d^3\mathbf{r}$ .

When propagation effects can be neglected, i.e., the speed of light is treated as infinite, in particular, when the electric dipole approximation is valid, the relations  $\text{div}(\mathbf{D}) = 0$  and  $\Delta V(\mathbf{r}) = 0$  remain valid.

### L. Comment on the Width $\Gamma$

In quantum calculations, one parameter is used to reduce the computation time, acting rather as a spectrum is smoothed by convolution with a function of finite width. More precisely, to a first approximation, this parameter serves

to give a finite width to the one-electron excitations (here 0.1 eV in Fig. 7.24). Figure 7.24 shows the spectra calculated with a parameter ten times smaller for cluster sizes  $N = 58, 138$  and  $1284$ . The fragmentation of the surface plasmon band of the cluster  $\text{Na}_{1284}$  is now clearly visible. Experimentally, except for very small sizes, this fragmentation phenomenon is rarely visible because many other effects (the same as those reducing the contrast of the shell effects in mass spectra, as discussed in Sects. 7.4.1 and 7.4.2) contribute to smoothing and broadening the absorption spectrum. These effects make it very difficult to produce reliable estimates of the intrinsic width  $\Gamma(N)$  of the plasmon band of a cluster at rest and the way it depends on the cluster size. For example, the widths observed in Fig. 7.23 are essentially dominated by temperature effects and shape fluctuations, since the technique of photoevaporation spectroscopy produces hot clusters. Furthermore, the temperature (and temperature-induced effects) are not necessarily the same for all clusters, since the multiphoton excitation is adjusted to each size in order to induce an appropriate evaporation sequence for analysis of the fragment distribution. Likewise, theoretical study of the size dependence of the intrinsic width of the plasmon band requires a significant reduction of the smoothing parameter. The width of the resonance bands in Fig. 7.24 is largely dominated by this artifact, despite a clearly visible average reduction, with the exception of the band associated with the size  $N = 138$ , which is much narrower. The kind of calculations carried out in these studies lead to a mean law of the type

$$\Gamma(N) = \Gamma_0 + \alpha/R_N, \quad (7.190)$$

where  $\Gamma_0$  and  $\alpha$  are two constants. The second term, induced by spatial confinement of the electrons, is introduced in the classical approach by including, in addition to the collisional mechanisms in the bulk solid phase (Drude parameter  $\Gamma$ ), a term corresponding to collision with the surface of the form  $Av_F/R_{N_e}$ , where  $v_F$  is the Fermi speed and  $A$  an adjustable phenomenological parameter which is very sensitive to the interface. This leads to a reduction in the mean free path of the electrons.

### M. Synthesis of Gold Clusters by Reducing with Citrate

Put 20 mg of hydrogen tetrachloroaurate trihydrate ( $\text{HAuCl}_4 \cdot 3\text{H}_2\text{O}$ ) in 190 mL of distilled water in a 500-mL three-necked flask, and heat to  $97^\circ\text{C}$  in a water bath, or over a Bunsen burner, stirring all the time. To allow for the reduced volume due to evaporation, a cooling column can be fitted to the flask. When the solution begins to boil, add 10 mL of a 1% by mass aqueous solution of sodium citrate. Continue to stir for ten or fifteen minutes. After this time, the solution will have become ruby red, indicating the presence of gold nanoparticles in the solution. The mean diameter of the nanoparticles is then 20 nm with standard deviation about 15%. The exact diameter can be determined by transmission electron imaging or granulometry. Still stirring,

allow the solution to return to room temperature. The stability of the solution is maintained by the citrate, which is a reducing agent, introduced in excess into the solution. In fact, the solution will remain stable for several weeks or even months.

## References

1. MacKenzie, M.J.K., Moore, A.J.W., and Nicholas, J.F.: *J. Phys. Chem. Solids* **23**, 185 (1962)
2. Jamet, M., et al.: *Phys. Rev. Lett.* **86**, 4676 (2001)
3. Kittel, C.: *Introduction to Solid State Physics*, 7th edn. Wiley, New York (1996). Gives definitions of the types of binding and the numerical values quoted for the Madelung constant
4. Joyes, P.: *Les agrégats inorganiques élémentaires*, EDP Sciences, Paris (1990). Provides details of this model
5. Conway, J.H., and Sloane, N.J.A.: *Sphere Packings, Lattices and Groups*, Comprehensive Studies in Mathematics, Springer Verlag (1991) p. 450
6. Van Hardeveld, R., and Hartog, F.: *Surf. Sci.* **15**, 189 (1969)
7. Flueli, M.: PhD thesis (no. 796) Lausanne EPFL, Switzerland (1989)
8. Sivardière, J.: *La symétrie en mathématiques, physique et chimie*, Grenoble Sciences, Presse Universitaire de Grenoble (1995). Provides the necessary topology
9. Buffat, P., and Borel, J.P.: *Phys. Rev. A* **13**, 2287 (1976)
10. Schumacher, E.: *Chimia* **42**, 357 (1988)
11. Halperin, W.P.: *Rev. Mod. Phys.* **58**, 533 (1986)
12. de Heer, W.A.: The physics of simple metal clusters: Experimental aspects and simple models, *Rev. Mod. Phys.* **65**, 611–675 (1993). Discusses quantum aspects
13. Brack, M.: The physics of simple metal clusters: Self-consistent jellium model and semiclassical approaches, *Rev. Mod. Phys.* **65**, 677–731 (1993). Discusses quantum aspects
14. Kresin, V.V.: Collective resonances and response properties of electrons in metal clusters, *Phys. Rep.* **220**, 1–52 (1992). Discusses quantum aspects
15. Kreibig, U., and Vollmer, M.: *Optical Properties of Metal Clusters*, Springer-Verlag, Berlin, Heidelberg (1995). Discusses classical and quantum aspects, as well as the collective optical properties of a particle ensemble
16. Bohren, C.F., and Huffman, D.R.: *Absorption and Scattering of Light by Small Particles*, Wiley Science Paperback Series, John Wiley, New York (1983). Discusses classical aspects for an isolated particle and an ensemble of particles
17. Knight, W.D., et al.: *Phys. Rev. Lett.* **52**, 2141 (1984)
18. Knight, W.D., Clemenger, K., de Heer, W.A., Saunders, W.A., Chou, M.Y., Cohen, M.L.: Electronic shell structure and abundances of sodium clusters, *Phys. Rev. Lett.* **52**, 2141–2143 (1984)
19. Bagueard, B., et al.: *J. Chem. Phys.* **100**, 754 (1994)
20. Pedersen, J., et al.: *Nature* **353**, 733 (1991)
21. Pellarin, M., et al.: *Phys. Rev. B* **52**, 16807 (1995)
22. Bréchnignac, C., et al.: *Phys. Rev. Lett.* **70**, 2036–2039 (1993)
23. Bréchnignac, C., et al.: *Phys. Rev. Lett.* **68**, 3916–3919 (1992)
24. Knight, W.D., et al.: *Phys. Rev. B* **31**, 2539–2540 (1985)

25. Palpant, B.: Doctoral thesis, University of Lyon I, France (1998)
26. Davidovits, P., and Faist, M.B.: J. Chem. Phys. **74**, 637 (1981)
27. Sun, Y., and Xia, Y.: Science **298**, December 2002, p. 2176
28. Kiely, C.J., Fink, J., Brust, M., Bethell, D., and Schiffrin, D.J.: Nature **396**, December 1998, p. 444
29. Haberland, H., Insepov, Z., and Moseler, M.: Molecular dynamic simulation of thin film growth by energetic cluster impact, Phys. Rev. B **51**, 11061–11067 (1995)
30. Jensen, P.: Growth of nanostructures by cluster deposition: Experiments and simple models, Rev. Mod. Phys. **71**, 1695 (1999)
31. Bardotti, L., Prével, B., Treilleux, M., Mélinon, P., and Perez, A.: Deposition of preformed gold clusters on HOPG and gold substrates: Influence of the substrate on the thin film morphology, Appl. Surf. Sci. **164**, 52–59 (2000)
32. Deltour, P., Barrat, J.L., and Jensen, P.: Fast diffusion of a Lennard-Jones cluster on a crystalline surface, Phys. Rev. Lett. **78**, 24, 4597–4600 (1997)
33. Bardotti, L., Prével, B., Jensen, P., Treilleux, M., Mélinon, P., Perez, A., Gierak, J., Faini G., and Maily, D.: Organizing nanoclusters on functionalized surfaces, Appl. Surf. Sci. **191**, 205–210 (2002)
34. Wood, D.M.: Phys. Rev. Lett. **46**, 749 (1981)
35. Landau, L.D., and Lifshitz, E.M.: *Electrodynamics of Continuous Media*, Pergamon Press, New York (1960)

## Fullerenes and Carbon Nanotubes

J.-P. Bourgoïn, A. Loiseau, and J.-F. Nierengarten

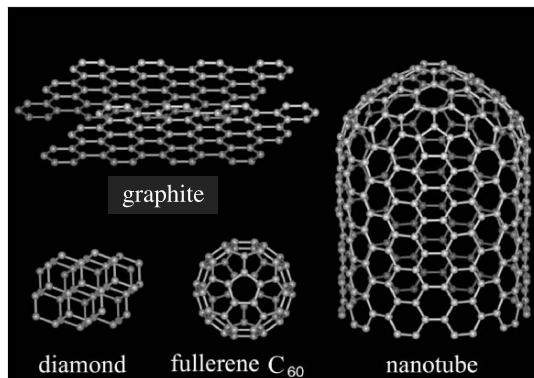
This chapter is a general introduction to fullerenes, carbon nanotubes, and related subjects, giving a snapshot of the state of the art ten years after their discovery. We show how these structures relate to other crystallised carbon structures and describe their specific distinguishing features, with a discussion of the various methods developed for observing and investigating them. We then review the main ways of synthesising these entities, and our current understanding of their formation mechanisms and properties.

### 8.1 Introduction

Fullerenes and carbon nanotubes still represent a new field of study, since they were only identified in 1985 [1] and they have only been studied since 1990 in the case of fullerenes [2] and 1991 in the case of carbon nanotubes [3]. These two discoveries cast a completely new light on carbon as an element, and the international scientific community reacted immediately, with such vigour that two associated fields of research were born, which soon established their importance in the world of science, both from a fundamental standpoint and in terms of applications.

To give an idea of the magnitude of the phenomenon, the subject is expanding exponentially if gauged by the number of publications: 2000 per year at present, having doubled between 1999 and 2002. This deep and long-lasting development is to a large extent due to the very special geometric characteristics of carbon nanotubes, with dimensions that are both nanometric and micrometric, leading to novel and versatile properties. For this reason, the related fields of research and applications are vast and ever more relevant to other branches of physics, chemistry and even biology. Nanotubes are thus particularly appropriate for scientific developments at the interface of these three disciplines.

The aim in this chapter is to attempt to explain what fullerenes and nanotubes are, how they are synthesised and studied, what properties they have,



**Fig. 8.1.** Structures of graphite, diamond, the  $C_{60}$  molecule and a carbon nanotube. Carbon atoms are represented by *grey balls* and chemical bonds between neighbouring atoms by *lines*. Adapted from [4]

and what the future prospects are, both in fundamental research and in applications.

## 8.2 Nanotubes and the Crystalline Forms of Carbon

### 8.2.1 Diamond and Graphite

In its natural state, carbon exists as a solid with two possible crystal forms familiar to everyone: graphite and diamond (see Fig. 8.1). Graphite is a crumbly black mineral, used for centuries as a writing material (Indian ink, pencil lead). In fact, previously called plumbago or black lead, its present name was only introduced in the eighteenth century from the Greek verb *graphein*, meaning ‘to write’.

In contrast, diamond is a transparent mineral, and extremely hard. It was only at the end of the eighteenth century that it was identified as a crystal form of carbon by Lavoisier and Tennant. In the diamond structure, each atom is connected to four neighbouring atoms arranged at the vertices of a regular tetrahedron by hybrid bonds of type  $sp^3$ . The tetrahedral symmetry signals a dense and isotropic solid. The distance between neighbouring atoms is 0.136 nm.

Graphite is layered or lamellar in structure, made up of parallel planes, each of which is a regular tiling of hexagons or honeycomb pattern. The carbon atoms are arranged at the vertices of the hexagons. Each atom is connected to three neighbouring atoms in the plane of hexagons by hybrid bonds of type  $sp^2$ , making an angle of  $120^\circ$  between them. The bonds in the plane are strong and characterised by an interatomic distance of 0.142 nm. However, an atom is only weakly connected to atoms in neighbouring planes. Indeed, the



distance between planes is 0.34 nm. This structure has a density equal to just one third of the density of diamond. Graphite is thus a highly anisotropic, quasi-2D solid, since the weakly connected planes slide very easily relative to one another. (For more details concerning graphite and diamond, and solid carbon in general, the reader is referred to [5].)

### 8.2.2 Discovery of Fullerenes

Observations of interstellar space by radioastronomers revealed the existence of chains of carbon atoms in certain stars, the red giants. While seeking to produce conditions similar to those obtaining in such stars, in order to produce the same molecules in the laboratory, H.W. Kroto, R.F. Curl and R.E. Smalley had no idea that they were about to make a discovery that would revolutionise our understanding of carbon.<sup>1</sup> Indeed, when examining the carbon clusters formed within a very hot plasma, obtained by vaporising graphite using a laser, they found molecules with a cage structure, exclusively made up of carbon atoms: the fullerenes [1].

After graphite and diamond, a third form of pure carbon was thus discovered. Whereas the two other forms are solids made up of an infinite atomic lattice, the fullerenes are well defined molecules. As can be seen from Fig. 8.1, the  $C_{60}$  molecule and carbon nanotubes manifest a clear family likeness with graphite. Indeed, the hexagons in the  $C_{60}$  molecule are the same as those in graphite. However, it would be several years before it became possible to study these compounds, by virtue of a method for synthesising them in macroscopic quantities using an electric arc between two graphite electrodes [2].

In this chapter we shall present this new family of molecules, and in particular, buckminsterfullerene, or  $C_{60}$ , which is the most commonly produced and studied.<sup>2</sup>

### 8.2.3 Discovery of Carbon Nanotubes

In 1991, S. Iijima [3] had the idea of examining under the electron microscope a byproduct of  $C_{60}$  synthesis using the electric arc method mentioned above. This byproduct occurred in the form of a hard, black, filamentary deposit. He discovered the nanotubes in this deposit, identifying them as tubular objects with nanometric diameter and micrometric length, closed at the ends and made of perfectly graphitic carbon (see Fig. 8.1).

In fact, the production of these nanotubes was not without precedent, as the discovery of the fullerenes had been in 1985. Back in 1960, R. Bacon [6] had already developed the electric arc technique for producing the carbon filaments, or whiskers, and there is a strong likelihood that nanotubes as we

<sup>1</sup> The discovery of fullerenes earned them the Nobel Prize for Chemistry in 1996.

<sup>2</sup> The name comes from the similarity between the structure of the  $C_{60}$  molecule and the domes created by the architect Buckminster Fuller.

understand them today had already been produced in those early experiments. However, the means of observation of the day would not have been adequate to their identification.

More recently, in the 1970s, methods were developed for fabricating carbon fibres using chemical vapour deposition, with or without the help of a metallic catalyst (CCVD or CVD, respectively). Although these fibres are generally a thousand times bigger than nanotubes and the level of graphitic structure far inferior, various electron microscope observations carried out by M. Endo and S. Oberlin [7] revealed that, in certain conditions, nanotubules or nanofibres would turn up, similar in many ways to the structures observed by S. Iijima. Although S. Iijima may not have actually discovered the nanotubes, he nevertheless realised the importance of these extraordinary nano-objects and drew attention to the remarkable significance of their dimensions.

## 8.3 Fullerenes

### 8.3.1 Structure of Fullerenes

Fullerenes are molecules with a cage structure containing  $2(10 + n)$  carbon atoms which form 12 pentagons and  $n$  hexagons. A hexagonal lattice forms a plane surface. Geometrically, the introduction of a pentagon into this lattice transforms the plane into an open cone with apex angle  $112^\circ$ . Introducing further pentagons, the plane can be closed off, transforming it into a shell. Euler's theorem tells us that 12 pentagons suffice to close the shell and arrive at a closed polyhedron. The smallest fullerene that can be imagined theoretically is  $C_{20}$ . Above  $C_{20}$ , any cluster made up of an even number of carbon atoms can form at least one fullerene-type structure. Increasing  $n$ , the number of fullerene isomers increases rapidly, from one for  $n = 0$  to more than 20 000 for  $n = 29$ .

Buckminsterfullerene  $C_{60}$  is the smallest stable fullerene. This compound has the shape of a truncated icosahedron, with a regular structure that corresponds to an Archimedean solid that can be inscribed in a sphere. It is in fact the exact replica of a football (see Fig. 8.2). Formed from 12 pentagons and 20 hexagons, with each pentagon surrounded by 5 hexagons,  $C_{60}$  is a highly symmetrical molecule in which all the carbon atoms are equivalent. It should be noted that there are two types of carbon-carbon bond in this molecule. Bonds at the join of two hexagons are 6-6 bonds, while those at the join of a hexagon and a pentagon are 5-6 bonds. Now the 6-6 bonds are shorter than the 5-6 bonds. Hence, the 6-6 bonds have the character of a double bond, whereas the 5-6 bonds have the character of a single bond. This localisation of the  $\pi$  electrons arises due to the pyramidalisation of the  $sp^2$  carbon atoms,

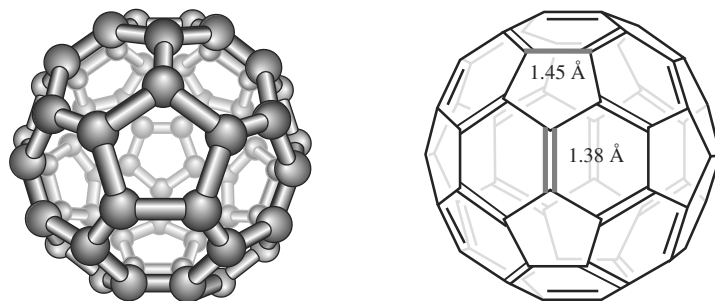


Fig. 8.2. Structure of  $C_{60}$

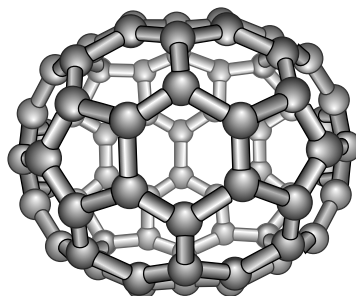


Fig. 8.3. Structure of  $C_{70}$

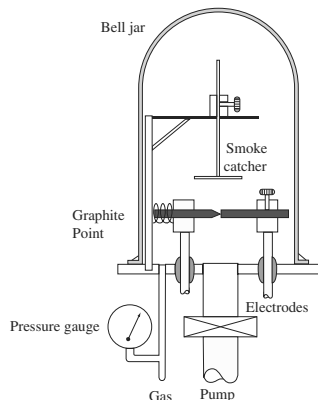
which in turn results from the fact that the spherical structure prevents full orbital overlap.<sup>3</sup>  $C_{60}$  is not therefore an aromatic molecule.

Buckminsterfullerene is the only isomer of  $C_{60}$ , and it is also the smallest fullerene obeying the isolated pentagon rule. This predicts that the fullerene-type structures in which all pentagons are separated from one another by hexagons are more stable than those in which there are two adjacent pentagons. The destabilisation arising from the presence of two such adjacent pentagons results essentially from a large tension in the carbon ring due to the presence of bond angles that are far removed from the standard value of  $120^\circ$ .

The second stable fullerene is  $C_{70}$ . Its structure also respects the isolated pentagon rule and it has an oval profile, rather like a rugby ball. At its poles,  $C_{70}$  has a structure similar to  $C_{60}$ , but it has an equatorial belt made up of a chain of hexagons (see Fig. 8.3).

Bearing in mind that only the fullerenes  $C_x$  respecting the isolated pentagon rule are stable, the magic numbers  $x$  are 60, 70, 72, 76, 78, 84, and so on. The number of theoretically possible isomers is one for  $C_{60}$ , one for  $C_{70}$ ,

<sup>3</sup> In benzene, the molecule is planar and the six carbon-carbon bonds have the same length due to the perfect delocalisation of the  $\pi$  electrons. This confers an even greater stability and specific properties on this aromatic molecule.



**Fig. 8.4.** Experimental setup for fullerene production [8]

one for  $C_{72}$ , one for  $C_{76}$ , five for  $C_{78}$ , 24 for  $C_{84}$ , and 46 for  $C_{90}$ . All these fullerenes, with the exception of  $C_{72}$ , are in fact obtained when graphite is evaporated in a helium atmosphere. A good proportion of these products have been isolated and characterised at the present time.

### 8.3.2 Production of Fullerenes

A technique for producing fullerenes in macroscopic quantities was developed in 1990 by Krätschmer and Huffman in 1990 [2]. It is based on the vaporisation of carbon in a helium atmosphere. The experimental setup comprises two graphite rods connected to copper electrodes (see Fig. 8.4). One of the rods is sharpened to a point and held in contact with the other by means of a spring. The whole thing is enclosed in a glass bell jar equipped with a pump to evacuate the air and a helium inlet. When an electric current passes through the graphite rods, heat produced by ohmic heating is dissipated mainly through the small point of contact between the two rods. The temperature there reaches  $2500\text{--}3000^\circ\text{C}$  and the graphite is vaporised as a plasma which cools on contact with the helium atmosphere to form a soot. This raw material comprises a mixture of soluble fullerenes  $C_n$  ( $n < 100$ ), so-called giant fullerenes  $C_n$  ( $n > 100$ ), nanotubes and amorphous carbon. Using suitable extraction techniques, the soluble fullerenes can be isolated from the soot.<sup>4</sup> Chromatographic methods are then used to separate the various fullerenes.

<sup>4</sup> This extract contains mainly  $C_{60}$  (roughly 60%) and  $C_{70}$  (roughly 20%), whilst the other fullerenes  $C_{76}$ ,  $C_{78}$ ,  $C_{82}$ ,  $C_{84}$ ,  $C_{90}$ ,  $C_{94}$ , and  $C_{96}$  are distinctly less abundant.

### 8.3.3 Physicochemical Properties of Buckminsterfullerene

The rest of the section will be devoted to  $C_{60}$ . Indeed, it is the most abundant of the fullerenes and certainly the one which has been most actively investigated. We should just note that most of the physicochemical characteristics of  $C_{60}$  carry over to its larger counterparts.

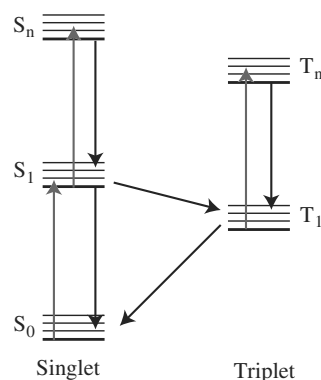
#### Solubility

Systematic studies have been carried out to determine the solubility of  $C_{60}$  in a wide range of organic solvents. It turns out to be insoluble in polar solvents such as acetone, the alcohols, tetrahydrofuran, diethyl ether, or dimethyl sulfoxide, and weakly soluble in hydrocarbons such as pentane, hexane or cyclohexane. The best solvents for  $C_{60}$  are the aromatic solvents, such as benzene (1.7 mg/mL), toluene (2.8 mg/mL), or 1-chloronaphthalene (51 mg/mL). It should be noted that solutions of  $C_{60}$  have a characteristic purple-magenta colour.

#### Photophysical Properties

Among the remarkable physicochemical properties of  $C_{60}$ , we note its optical nonlinearity and in particular a useful optical limitation effect due to its absorption characteristics [9]. Indeed,  $C_{60}$  absorbs visible light only weakly if the molecule is illuminated by low intensity light. However, when the light intensity increases, the level of absorption increases significantly. In other words, the absorption effect is nonlinear. This phenomenon can be used to protect an optical sensor, e.g., a camera or the human eye, against laser damage, without in any way impeding its use in low light conditions such as ordinary daylight.

The  $C_{60}$  molecule exhibits a high degree of symmetry. A theoretical description of the optical properties of an isolated molecule is possible as a result



**Fig. 8.5.** Five-level model explaining the photophysical properties of  $C_{60}$

of this symmetry, thereby explaining the experimentally observed nonlinear absorption. Theory predicts that the molecule will absorb much more efficiently when in an excited state, as compared to the ground state. Since the number of excited molecules in a sample depends directly on the incident light intensity, this leads to higher absorption for greater incident intensity. This phenomenon is known as inverse saturable absorption (ISA).

Several studies have been made of the nonlinear absorption by  $C_{60}$ , generally in a toluene solution. The results can be interpreted using a five-level model in the ISA framework (see Fig. 8.5). A large number of vibronic states of the molecule are associated with each electronic level and this considerably broadens the range of frequencies absorbed by the molecule. The vibronic states associated with electronic states are close in energy and partially overlapping. In this way, a simplified representation can be used to describe the optical properties. In this picture, the ground state of the molecule and its vibronic levels are associated with a level  $S_0$ . All the first group of excited singlet states are grouped together in a level  $S_1$ . Note that the transition  $S_0 \rightarrow S_1$  is forbidden for symmetry reasons and the associated absorption is thus low. The other excited singlet states of higher energy at the origin of allowed transitions from the ground state are associated with a level  $S_n$ . (The first allowed transition is in the blue-near UV range and a high level of absorption is observed in this part of the linear absorption spectrum.) The triplet states can also be distributed in groups  $T_1$  and  $T_n$ . In terms of nonlinear absorption, it turns out that certain transitions from the state  $S_1$  to the states  $S_n$  are allowed in the visible range of the spectrum. The same is true for transitions between the states making up the levels  $T_1$  and  $T_n$ . For this reason, a population in an excited state, either singlet or triplet, will absorb much more efficiently in the visible than a population in the ground state. If the  $C_{60}$  is illuminated with low light intensity, the population of molecules in the ground state is much bigger than the population in excited states, for the low intensity generates few transitions which could populate the excited level. At higher light intensities, the population in the first excited state becomes non-negligible and absorption increases with the population in this excited state, and hence with the incident intensity.

The absorption induced by a laser pulse can last for several microseconds. To limit long pulses, the existence of triplet states is of great importance, for an effective limitation requires an excited population with a lifetime that is at least comparable with the pulse width. Otherwise, an equilibrium is reached between  $S_0$ - $S_1$  absorption and  $S_1$ - $S_0$  relaxation during a long pulse. Then the mean excited population during the pulse will be small and hence the limitation too. A key parameter in the case of long pulses is the proportion of molecules undergoing intersystem crossing and accumulating in the level  $T_1$ , rather than decaying directly. The higher the proportion, the more effective the limitation for long pulses which require the presence of a triplet population with long enough lifetime. From this point of view,  $C_{60}$  is a good candidate

**Table 8.1.** Reduction potentials obtained at  $-10^{\circ}\text{C}$  in a  $\text{CH}_3/\text{toluene}$  mixture. Values are obtained in volts (vs.  $\text{Fc}^+/\text{Fc}$ ) for a scan rate of  $100\text{ mV s}^{-1}$ 

Single-electron pair	Reduction potential [V]
$\text{C}_{60}/\text{C}_{60}^-$	-0.98
$\text{C}_{60}^-/\text{C}_{60}^{2-}$	-1.37
$\text{C}_{60}^{2-}/\text{C}_{60}^{3-}$	-1.87
$\text{C}_{60}^{3-}/\text{C}_{60}^{4-}$	-2.35
$\text{C}_{60}^{4-}/\text{C}_{60}^{5-}$	-2.85
$\text{C}_{60}^{5-}/\text{C}_{60}^{6-}$	-3.26

because, in a toluene solution, intersystem crossing occurs with efficiency close to unity.

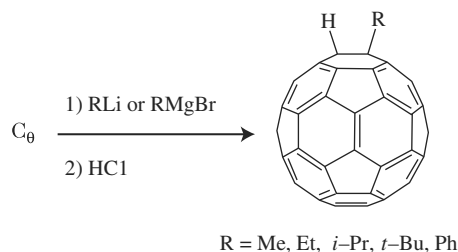
### Electrochemical Properties

Theoretical calculations predicted that the  $\text{C}_{60}$  molecule would have a rather low energy, triply degenerate unoccupied molecular orbital. Put more simply, this molecule was expected to behave as an electron acceptor and could in principle accept up to 6 electrons to form a hexa-anion  $\text{C}_{60}^{6-}$ . As soon as pure samples of  $\text{C}_{60}$  had been obtained, the electrochemical properties were studied and the theoretical predictions checked experimentally. It was indeed shown that  $\text{C}_{60}$  could accept up to 6 electrons by six successive single-electron reductions (see Table 8.1) [10]. It should be noted that these reductions are reversible processes, and also that the anions thus obtained remain stable for several days at low temperatures. It should also be stressed that, although  $\text{C}_{60}$  is relatively easy to reduce (at least for the first reduction leading to the anion  $\text{C}_{60}^-$ ), it is rather difficult to oxidise.

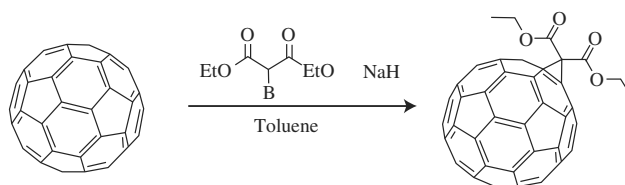
These are important results, offering a first clue as to the chemical reactivity of  $\text{C}_{60}$ . Indeed, the electronegativity of  $\text{C}_{60}$  suggests that this molecule will have an electrophilic tendency, hosting addition reactions in the presence of nucleophilic reagents.

### Chemical Properties

Ever since a process was perfected for synthesising macroscopic quantities of  $\text{C}_{60}$  in 1990, its chemical reactivity has been the subject of a great many studies. Reactions were found which could graft groups of molecules onto the surface of  $\text{C}_{60}$  and a large number of derivatives of  $\text{C}_{60}$  have now been produced. The chemical modification of  $\text{C}_{60}$  has one aim in particular, namely to increase its solubility, whereupon the corresponding derivatives become much easier to manipulate. Furthermore, many molecular groups with specific functions have been successfully grafted onto  $\text{C}_{60}$  to yield molecules with novel properties. This is not the place to summarise all the work done in this



**Fig. 8.6.** Nucleophilic addition of an organolithic or organomagnesium compound to  $C_{60}$  followed by acid hydrolysis



**Fig. 8.7.** Cyclopropanation of  $C_{60}$

area and the reader interested in obtaining further detail is referred to the literature [8, 11]. We shall limit ourselves here to giving a few examples of chemical reactions wherein a molecular group is grafted onto  $C_{60}$ .

$C_{60}$  has a similar chemical reactivity to an electron-deficient olefin. Hence,  $C_{60}$  is a good electrophilic reagent and can host nucleophilic addition reactions. For example,  $C_{60}$  reacts with various nucleophilic derivatives such as organolithic or organomagnesium compounds. The corresponding salts  $C_{60}RM$  generated rapidly by addition of the organometallic compound to  $C_{60}$  can then be protonated in an acid medium to yield the corresponding hydroalkylated or hydroarylated derivatives (see Fig. 8.6).

It has been possible to carry out a cyclopropanation reaction of  $C_{60}$  by means of an addition/elimination mechanism (see Fig. 8.7). This involves the reaction of a stabilised  $\alpha$ -halocarbanion with  $C_{60}$ . The mechanism here can be divided into two steps. First the  $C_{60}$  reacts with the nucleophilic reagent  $\alpha$ -halocarbanion, itself generated by a reaction between the base NaH and  $\alpha$ -bromo-malonate. Then in the second step the cyclopropanation product is generated by an intramolecular nucleophilic substitution.

$C_{60}$  can also play the role of dienophile or 1,3-dipolarophile (see Fig. 8.8). Hence, cycloadditions of all types leading to the synthesis of cyclic derivatives are possible, i.e., carbene- or nitrene-insertion [2 + 1] cycloadditions, Diels–Alder-type [4 + 2] cycloadditions, thermally or photochemically induced [2 + 2] cycloadditions, or 1,3-dipolar [3 + 2] cycloadditions.



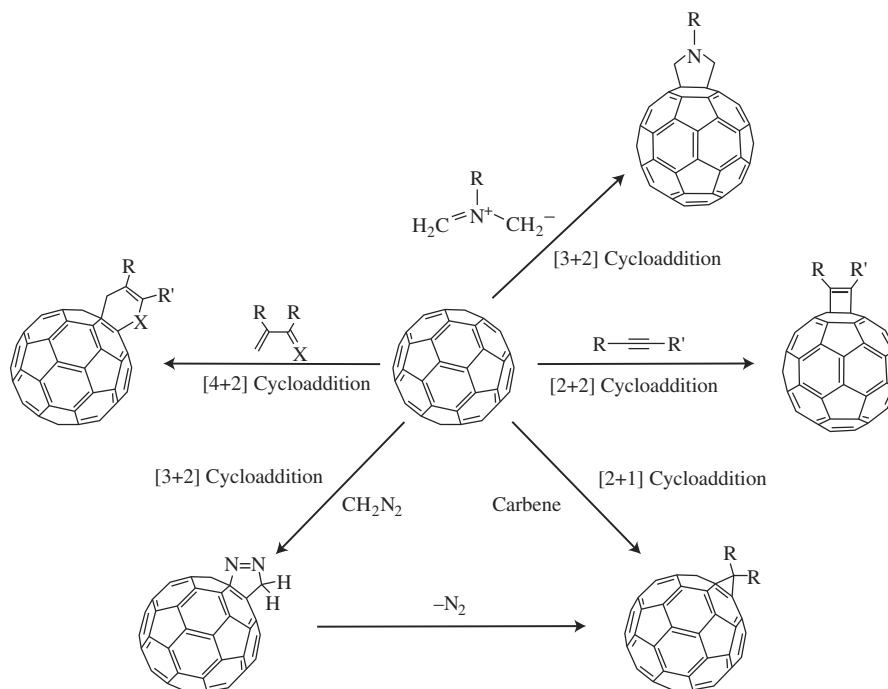


Fig. 8.8. Different cycloaddition reactions involving  $C_{60}$

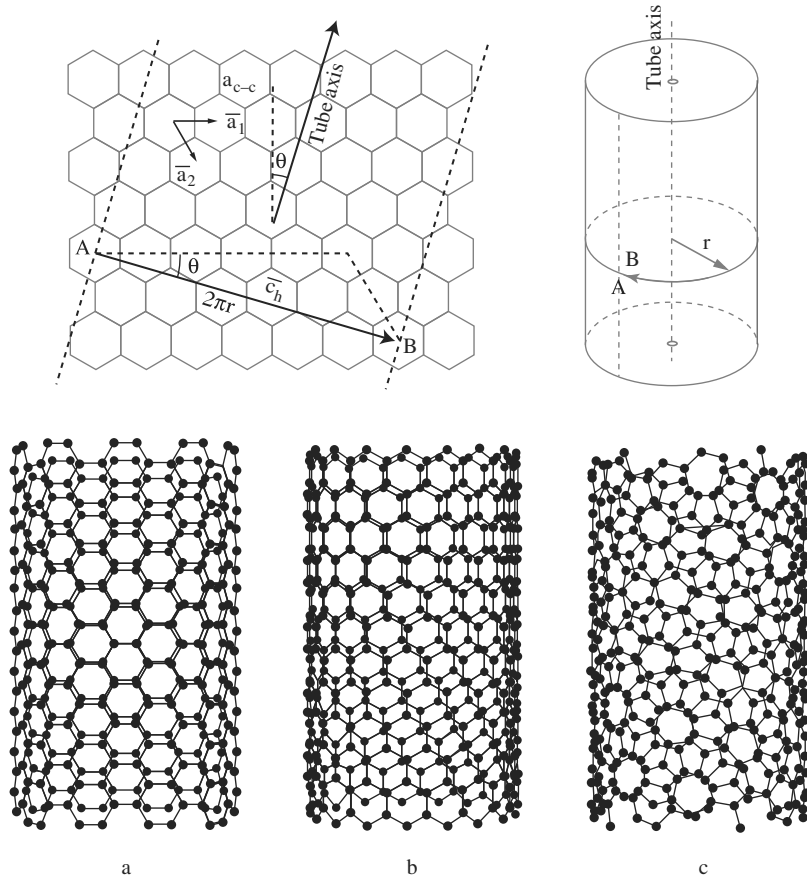
## 8.4 Carbon Nanotubes

### 8.4.1 Crystal Structure of Nanotubes

The most striking feature of these objects, as mentioned earlier, is their dimensions, which give them their name. They can be several microns long, but their diameter is of nanometric order, typically in the range 1–10 nm. A diameter of 1 nm is of the same order as the double helix in the DNA molecule, indicating the molecular nature of these nanotubes.

The second striking feature of nanotubes is the crystal structure of the carbon making them up, compared with graphite and diamond, shown schematically in Fig. 8.1.

As can be seen from Fig. 8.1, carbon nanotubes are clearly related to graphite by their crystal structure. The atomic structure of the nanotube is in fact obtained by taking a sheet of hexagons from the graphite structure, known as graphene, and rolling it up to form a cylinder. This cylinder is open at the ends and cannot be closed without considerably distorting the hexagons and hence the  $sp^2$  bonds. Closing the tubes requires the introduction of topological defects into the plane hexagonal lattice, thereby curving this plane. As we have seen for the fullerenes, curvature is achieved by introducing pentagons. Each end of a nanotube corresponds to the closure of a half-space



**Fig. 8.9.** Construction of a nanotube by rolling up a graphene sheet. The three tubes in the *lower* part of the figure result from rolling the sheet up at different angles  $\theta$ . (a)  $\theta = 30^\circ$ , armchair configuration. (b)  $\theta = 0^\circ$ , zigzag configuration. (c)  $\theta \neq 0^\circ$  or  $30^\circ$ , chiral configuration

which is achieved by introducing 6 pentagons into the hexagonal lattice. The topology of the end depends on the distribution of these pentagons. A regular distribution defines a hemispherical end (the case illustrated in Fig. 8.1), whilst in the general case, one obtains a conically-shaped tip, as we shall see below.

To give a complete description of the nanotube structure, we must examine the way the graphene sheet is rolled up. As shown in Fig. 8.9, this operation amounts to superposing two hexagons A and B of the lattice and the result depends entirely and uniquely on the choice of these two hexagons. This choice fixes the diameter of the nanotube and an angle known as the chiral angle or helicity, which specifies the direction in which the sheet is rolled up. By

choosing a reference direction as determined by one side of a hexagon, we define the chiral angle  $\theta$  as the angle between the axis of the cylinder and this reference direction. This angle varies between  $0$  and  $30^\circ$ , given the symmetry of the hexagonal lattice, and allows a complete classification of all possible configurations into three categories called the armchair, zigzag and chiral configurations. Zigzag and armchair tubes have chiral angle equal to  $0^\circ$  and  $30^\circ$ , respectively. Their names refer to the arrangement of carbon atoms on the rim of an open tube (see Fig. 8.9). In these two types of tube, the hexagons in the upper part of the tube have the same orientation with respect to the axis as those in the lower part. Such tubes are said to be achiral. This property is not satisfied in tubes with chiral angle  $\theta$  different from  $0^\circ$  or  $30^\circ$ , which belong to the third category. The rows of hexagons in the upper region make an angle  $2\theta$  with the rows of hexagons in the lower region and they define an Archimedean screw when rolled up.

This can be stated formally by reiterating that the structure of carbon nanotubes is deduced from the structure of a graphene sheet which has been rolled up to form a cylinder and formulating as follows. The chiral vector or roll-up vector  $\mathbf{C} = n\mathbf{a}_1 + m\mathbf{a}_2$ , with  $n, m \in \mathbb{N}$ , characterises the nanotube as shown in Fig. 8.9. The figure illustrates the three types of nanotube: the armchair tube with indices  $(n, n)$ , the zigzag tube with indices  $(n, 0)$ , and the chiral tube with indices  $(n, m)$ . As we shall see below, the electronic properties of nanotubes are directly dependent on the chiral vector.

The diameter  $d$  of the tube, the chiral angle  $\theta$ , and the vector  $\mathbf{T}$  defining the unit cell are given by

$$d_{\text{NT}} = \frac{1}{\pi} \|\mathbf{C}\| = \frac{\sqrt{3}}{\pi} a_0 \sqrt{(m^2 + nm + n^2)},$$

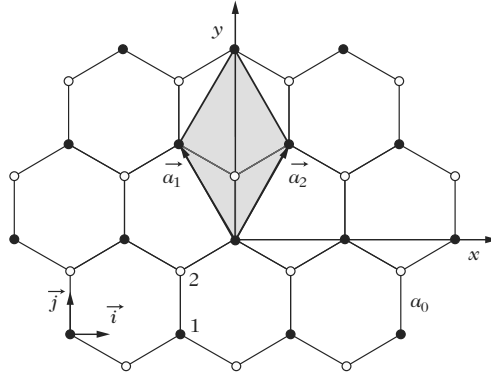
$$\theta = \arctan \frac{\sqrt{3}m}{m + 2n},$$

$$\mathbf{T} = \frac{(2m + n)\mathbf{a}_1 - (2n + m)\mathbf{a}_2}{d_r},$$

where

$$d_r = \begin{cases} \text{GCD}(m, n) & \text{if } (m - n) \text{ not multiple of } 3 \times \text{GCD}(m, n), \\ 3 \times \text{GCD}(m, n) & \text{if } (m - n) \text{ multiple of } 3 \times \text{GCD}(m, n), \end{cases}$$

and GCD denotes the greatest common divisor. The full set of ingredients defining a nanotube can be stated as follows: the nanotube has a structure derived from graphite in which a simple curvature has been introduced along with several topological defects, and to which a one-dimensional aspect and molecular size have been attributed. These ingredients make the nanotube a quite unique object, with an equally unique cocktail of extraordinary properties, as we shall see below.



**Fig. 8.10.** Graphene lattice. The vectors  $\mathbf{a}_1$  and  $\mathbf{a}_2$  are the basis vectors of the lattice, while  $\mathbf{i}$  and  $\mathbf{j}$  form an orthonormal basis for the plane.  $a_0$  represents the distance between two nearest-neighbour carbon atoms ( $a_0 = 0.136$  nm)

#### 8.4.2 Electronic Structure of Carbon Nanotubes

As we have just seen, the crystal structure of carbon nanotubes follows from the crystal structure of a single sheet of graphite, known as graphene. We shall begin by calculating the electronic band structure of graphene, and then deduce the same for the nanotubes. The reader will find a summary of the structural and electronic characteristics of nanotubes in Table 8.2 at the end of the chapter.

#### Graphene Structure

Graphene is made up of carbon atoms arranged in a hexagonal lattice as shown schematically in Fig. 8.10. This lattice is described by the two basis vectors

$$\mathbf{a}_1 = -\frac{\sqrt{3}}{2}a_0\mathbf{i} + \frac{3}{2}a_0\mathbf{j} \quad \text{and} \quad \mathbf{a}_2 = \frac{\sqrt{3}}{2}a_0\mathbf{i} + \frac{3}{2}a_0\mathbf{j}.$$

The unit cell contains two carbon atoms corresponding to the positions marked 1 and 2 in Fig. 8.10.

Each carbon atom has chemical bonds with three neighbouring atoms. There are two types of bond:  $\sigma$  bonds involving electrons in the  $2s$ ,  $2p_x$  and  $2p_y$  orbitals of carbon, and  $\pi$  bonds involving electrons in the  $2p_z$  orbitals.

The electronic band structure of graphene can be calculated using the tight-binding approximation, also known as the LCAO method (linear combination of atomic orbitals).<sup>5</sup> We consider only the electrons in the  $2p_z$  orbitals

<sup>5</sup> The reader is referred to the book by C. Kittel [49] for the general theory of band structures. Note also the Bloch theorem: the stationary wave functions (solutions

which give rise to the  $\pi$  band and account for the transport properties of graphene.

To calculate the electronic band structure of graphene, we consider the Bloch functions corresponding to the sublattices of carbon atoms 1 and 2:

$$\chi_1 = \frac{1}{\sqrt{N}} \sum_n e^{i\mathbf{k}\cdot\mathbf{R}_{1n}} \varphi_1(\mathbf{r} - \mathbf{R}_{1n}),$$

$$\chi_2 = \frac{1}{\sqrt{N}} \sum_n e^{i\mathbf{k}\cdot\mathbf{R}_{2n}} \varphi_2(\mathbf{r} - \mathbf{R}_{2n}),$$

where  $\mathbf{k}$  is the wave vector,  $N$  is the arbitrarily large number of unit cells considered,  $\varphi_1(\mathbf{r})$  and  $\varphi_2(\mathbf{r})$  represent the  $p_z$  orbitals of carbon atoms 1 and 2, respectively,  $n$  denotes the unit cell  $n$  found from the cell shown in Fig. 8.9 by translating by a lattice vector, and  $\mathbf{R}_{1n}$  and  $\mathbf{R}_{2n}$  correspond to the positions of carbon atoms 1 and 2, respectively, in cell  $n$ . We seek a solution to the Schrödinger equation in the form of a linear combination of these Bloch functions, i.e.,  $\chi = a\chi_1 + b\chi_2$ .

Let  $H$  be the Hamiltonian for an electron in the atomic potential given by the atoms in the graphene lattice. In order to calculate the terms  $\langle\chi|H|\chi\rangle$ , one must first find the terms  $H_{ij} = \langle\chi_i|H|\chi_j\rangle$ . In the tight-binding formalism, only the following terms are considered:

- $\langle\varphi_a|h|\varphi_a\rangle = \alpha$ , where  $h$  is the mean field Hamiltonian and  $\varphi_a$  is the  $2p_z$  atomic orbital of a carbon atom  $a$ ,
- $\langle\varphi_a|h|\varphi_b\rangle = \beta$ , if the carbon atoms  $a$  and  $b$  are nearest neighbours,
- $\langle\varphi_a|\varphi_b\rangle = \delta_{ab}$ .

Without loss of generality, we shall set  $\alpha = 0$ , since this corresponds to a simple shift in the zero energy.

It follows that

$$H_{11} = H_{22} = \langle\chi_1|H|\chi_1\rangle = \alpha = 0,$$

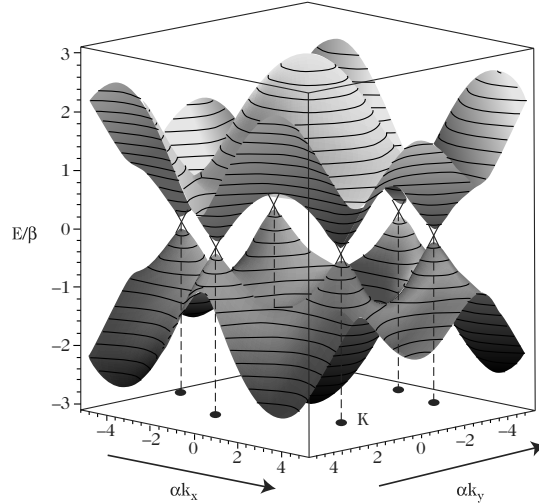
$$H_{12} = H_{21}^* = \langle\chi_1|H|\chi_2\rangle.$$

The integral can be calculated by observing that each atom is surrounded by three neighbours. For an atom of type 1, neighbours are of type 2 and their relative coordinates are  $(1/3, 1/3)$ ,  $(1/3, -2/3)$ ,  $(-2/3, 1/3)$ . Hence,

$$H_{12} = \beta \left\{ \exp \left[ i\mathbf{k}\cdot \left( \frac{\mathbf{a}_1}{3} + \frac{\mathbf{a}_2}{3} \right) \right] + \exp \left[ i\mathbf{k}\cdot \left( \frac{\mathbf{a}_1}{3} - \frac{2\mathbf{a}_2}{3} \right) \right] + \exp \left[ i\mathbf{k}\cdot \left( \frac{-2\mathbf{a}_1}{3} + \frac{\mathbf{a}_2}{3} \right) \right] \right\}.$$

---

of the Schrödinger equation) of an electron in a periodic crystal can be written as a product of a plane wave  $\exp(i\mathbf{k}\cdot\mathbf{r})$  and a function  $u_k(\mathbf{r})$  with the periodicity of the lattice.



**Fig. 8.11.** Electronic band structure of graphene

Projecting  $H|\chi\rangle = e(\mathbf{k})|\chi\rangle$  onto  $\langle\chi_1|$ , we obtain  $aH_{11} + bH_{12} = ae(\mathbf{k})$ . Likewise, the projection onto  $\langle\chi_2|$  yields  $aH_{21} + bH_{22} = be(\mathbf{k})$ . The only nonzero solution for the coefficients  $a$  and  $b$  is given by the secular determinant, viz.,

$$\begin{vmatrix} -e(\mathbf{k}) & H_{12} \\ H_{21} & -e(\mathbf{k}) \end{vmatrix} = 0.$$

The solution of this equation gives the dispersion relation

$$e(\mathbf{k}) = \pm\beta\sqrt{3 + 2\cos\mathbf{k}\cdot(\mathbf{a}_1 - \mathbf{a}_2) + 2\cos\mathbf{k}\cdot\mathbf{a}_1 + 2\cos\mathbf{k}\cdot\mathbf{a}_2},$$

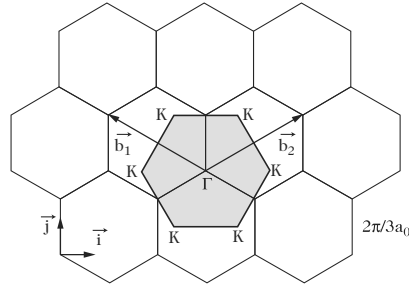
which can be rewritten in the form

$$e(k_x, k_y) = \pm\beta\sqrt{1 + 4\cos\left(\frac{\sqrt{3}a}{2}k_y\right)\cos\left(\frac{a}{2}k_x\right) + 4\cos^2\left(\frac{a}{2}k_x\right)},$$

where  $a = a_0\sqrt{3}$ . The corresponding electronic band structure is shown in Fig. 8.11. The valence band corresponding to negative energies (since we took  $\alpha = 0$ ) is totally filled because the number of electrons is equal to the number of orbitals. The Fermi level corresponds to zero energy.

Note that, at six points, the valence band and the conduction band touch one another. These points K are also indicated in Fig. 8.12, which shows the first Brillouin zone of graphene in the reciprocal lattice.<sup>6</sup>

<sup>6</sup> The first Brillouin zone is the smallest volume entirely enclosed between the mediating planes of the reciprocal lattice vectors drawn from the origin marked  $\Gamma$  in Fig. 8.12 [49].



**Fig. 8.12.** Reciprocal lattice and first Brillouin zone of graphene (*shaded*). The vectors  $\mathbf{b}_1$  and  $\mathbf{b}_2$  are the basis vectors of the lattice

The basis vectors  $\mathbf{b}_1$  and  $\mathbf{b}_2$  of the reciprocal lattice are defined by  $\mathbf{b}_i \cdot \mathbf{a}_j = 2\pi\delta_{ij}$ , or

$$\mathbf{b}_1 = \frac{4\pi}{3a_0} \left( -\frac{\sqrt{3}}{2}\mathbf{i} + \frac{1}{2}\mathbf{j} \right), \quad \mathbf{b}_2 = \frac{4\pi}{3a_0} \left( \frac{\sqrt{3}}{2}\mathbf{i} + \frac{1}{2}\mathbf{j} \right).$$

It is due to the existence of these points K where the valence and conduction bands meet that graphite is a semi-metal.

### Electronic Structure of Carbon Nanotubes

The electronic structure of carbon nanotubes is determined from that of graphene by noting that the periodic boundary condition imposes the following quantisation rule:

$$\mathbf{C} \cdot \mathbf{k} = 2\pi q, \quad \text{where } q \in \mathbb{Z}.$$

#### *Armchair Nanotubes*

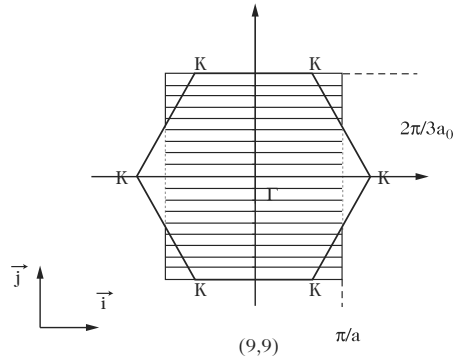
Let us apply this to the armchair nanotubes, for which  $\mathbf{C} = n(\mathbf{a}_1 + \mathbf{a}_2)$  and  $\mathbf{T} = \mathbf{a}_2 - \mathbf{a}_1$ . The repeat unit of the armchair nanotube contains  $2n$  unit cells of the graphene lattice and thus involves  $2 \times 2 \times n$  valence electrons. The vector  $\mathbf{T} = \sqrt{3}a_0\mathbf{i} = a\mathbf{i}$  leads us to define the Brillouin zone of the armchair nanotube by

$$-\pi/a < k_x < \pi/a.$$

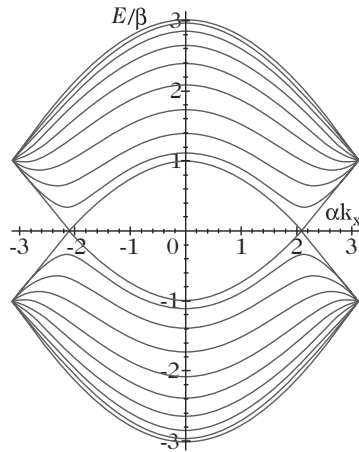
For armchair nanotubes, expanding  $\mathbf{C}$  in terms of the basis  $(\mathbf{i}, \mathbf{j})$ , the quantisation relation implies that the wave vector is quantised along  $y$  and equal to

$$k_y = \frac{q2\pi}{3na_0}, \quad q \in \mathbb{Z},$$

whence



**Fig. 8.13.** Quantisation conditions for a (9,9) armchair nanotube on the first Brillouin zone of graphene



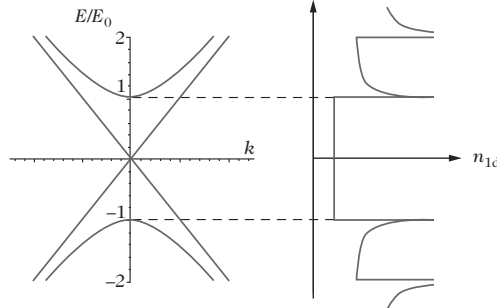
**Fig. 8.14.** Electronic band structure of a (9,9) armchair nanotube

$$e(k_x, k_y) = \pm \beta \sqrt{1 + 4 \cos\left(\frac{q\pi}{n}\right) \cos\left(\frac{a}{2}k_x\right) + 4 \cos^2\left(\frac{a}{2}k_x\right)}.$$

This relation shows that the electronic band structure of the armchair nanotube can be deduced from that of graphene by cutting the latter by planes defined by the quantised values of  $k_y$ . These planes are represented in Fig. 8.13. The electronic band structure of the armchair nanotube is deduced for  $1 \leq q \leq n$  (see Fig. 8.14).

The  $n$  values of  $q$  give rise to  $n$  corresponding subbands. The fold over the first Brillouin zone gives rise to an  $(n + 1)$ th dispersion relation corresponding to  $q = 0$ . (This is the fold in the band  $q = n$  of the zone  $-\pi/a < k_x < \pi/a$ .) This electronic band structure thus includes 10 subbands with positive energies (the Fermi energy being zero, as mentioned above) and ten subbands





**Fig. 8.15.** Close-up view of the Fermi region showing the band structure of the armchair nanotube, plotted in reduced units  $E/E_0$ , where  $E_0 = \beta\pi/n$  with  $\beta$  the interaction term between nearest neighbours and  $n$  the index defining the  $(n, n)$  armchair nanotubes and associated density of states

with negative energies. Apart from the bands corresponding to  $q = 0$  and  $n$ , each of these bands is doubly degenerate. This is a consequence of the fact that there are  $2n$  graphite unit cells in the repeat unit of the nanotube.

The fact that there are two bands at the Fermi level (for  $k_x = \pm 2\pi/3a$ ) shows that the armchair nanotube is metallic. This shows up clearly in the density of states  $g(k) = (a/2\pi)\partial k/\partial e(k)$  plotted in Fig. 8.15, which is nonzero at the Fermi level.

#### Zigzag Nanotubes

The zigzag nanotubes are defined by  $\mathbf{C} = n\mathbf{a}_1$  and  $\mathbf{T} = \mathbf{a}_1 - 2\mathbf{a}_2$ . We can understand them by the same kind of analysis as for the armchair nanotubes. The quantisation condition leads to

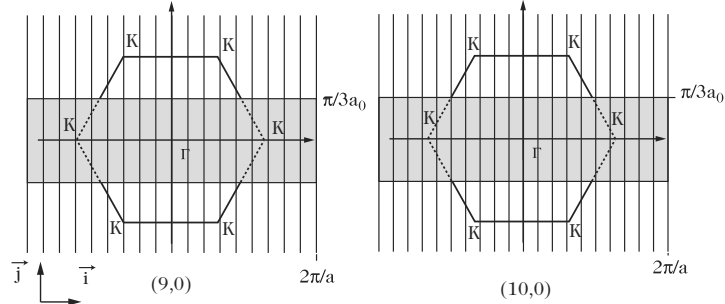
$$-\frac{1}{2}k_x + \frac{\sqrt{3}}{2}k_y = q\frac{2\pi}{\sqrt{3}na_0}, \quad q \in \mathbb{Z}.$$

Relative to the frames of reference in Figs. 8.9 and 8.10, an equivalent condition given the symmetry of the problem is

$$k_x = \frac{q2\pi}{\sqrt{3}na_0}, \quad q \in \mathbb{Z}.$$

The Brillouin zone of the zigzag nanotube is sketched in Fig. 8.16 for the two zigzag nanotubes specified by indices  $(9,0)$  and  $(10,0)$ . It is defined by  $-\pi/\sqrt{3}a < k_y < \pi/\sqrt{3}a$ . Using the quantisation condition, the dispersion relation is

$$e(k_x, k_y) = \pm\beta\sqrt{1 + 4\cos\left(\frac{\sqrt{3}a}{2}k_y\right)\cos\left(\frac{q\pi}{n}\right) + 4\cos^2\left(\frac{q\pi}{n}\right)},$$



**Fig. 8.16.** Planes defining the allowed values of  $k$  for (9,0) and (10,0) zigzag nanotubes. The *shaded regions* are the extended Brillouin zones of the nanotubes

and the band structures for the (9,0) and (10,0) nanotubes have been drawn in Fig. 8.17.

As for the armchair nanotubes, the electronic band structure folded over the first Brillouin zone gives rise to an  $(n+1)$ th subband corresponding to  $q = 0$ . (This is the fold in the band  $q = n$  of the zone  $-3\pi/\sqrt{3}a < k_y < -\pi/\sqrt{3}a$ .) This electronic band structure thus contains  $n+1$  subbands with positive energies (the Fermi level being zero, as mentioned above), and  $n+1$  subbands with negative energies. Apart from the bands corresponding to  $q = 0$  and  $n$ , each of these bands is doubly degenerate.

It can be seen from Figs. 8.16 and 8.17 that the (9,0) zigzag nanotubes are metallic, one of the subbands containing a graphene K point, whereas the (10,0) nanotubes are semiconductors, with a forbidden energy gap, which corresponds to the fact that none of the subbands contains any K point. This result can be generalised as follows: the  $(n,0)$  zigzag nanotubes are metallic if and only if  $n$  is an integer multiple of 3.

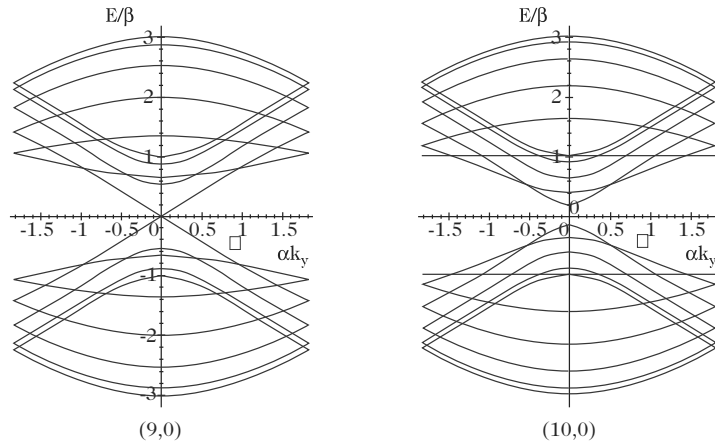
#### General Case

The general case of a nanotube with chiral vector  $\mathbf{C} = n\mathbf{a}_1 + m\mathbf{a}_2$  can be treated in the same way as the two others. The quantisation condition  $\mathbf{C} \cdot \mathbf{k} = 2\pi q$ , with  $q \in \mathbb{Z}$ , is used to define the planes (in the space of the vectors  $\mathbf{k}$  or the reciprocal space<sup>7</sup>) perpendicular to the vector  $\mathbf{C}$  corresponding to the allowed wave vectors  $\mathbf{k}$ . The dispersion relations corresponding to the subbands are then deduced from the general dispersion relation for graphene:

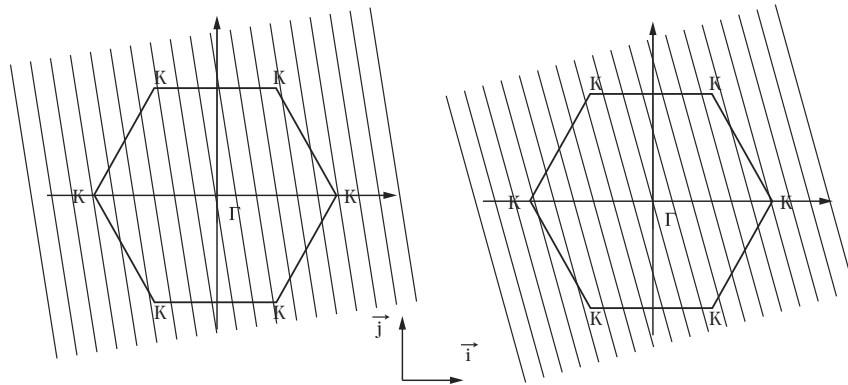
$$e(k_x, k_y) = \pm\beta \sqrt{1 + 4 \cos\left(\frac{\sqrt{3}a}{2}k_y\right) \cos\left(\frac{a}{2}k_x\right) + 4 \cos^2\left(\frac{a}{2}k_x\right)},$$

with  $k_x$  and  $k_y$  the components of the vector  $\mathbf{k}$  obeying the quantisation condition. As for the armchair and zigzag nanotubes, this amounts to segmenting

<sup>7</sup> The reciprocal lattice defined above is located in the reciprocal space.



**Fig. 8.17.** Electronic band structure of the (9,0) and (10,0) zigzag nanotubes on the first Brillouin zone

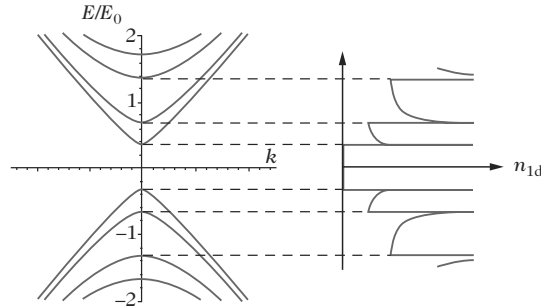


**Fig. 8.18.** *Left:* Metallic chiral nanotube. *Right:* Semiconducting chiral nanotube

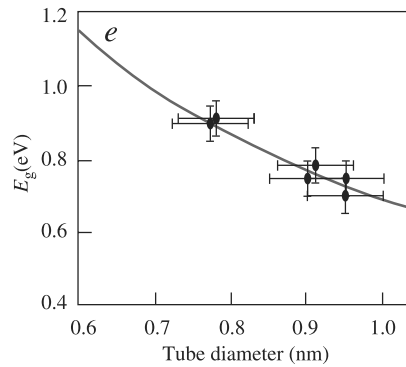
the Brillouin zone of graphene by straight lines perpendicular to  $\mathbf{C}$  and with spacing  $2\pi/|\mathbf{C}|$ .

Figure 8.18 shows two examples which differ by the fact that the electronic band structure of the nanotube contains or does not contain a graphene K point. It can be shown that this happens when the condition  $n - m = 3q$ ,  $q \in \mathbb{Z}$ , is satisfied. In this case the nanotube will be metallic, since this corresponds to the existence of two bands at the Fermi level, as in the case of the armchair nanotube. The electronic band structure and the density of states correspond in this case to those depicted in Fig. 8.15.

In the case  $n - m \neq 3q$ ,  $q \in \mathbb{Z}$ , the nanotube will be a semiconductor and will have a forbidden energy gap  $E_g$  which is inversely proportional to the diameter  $d_{NT}$  of the nanotube, viz.,  $E_g = 2\beta a_0/d_{NT}$ . Figure 8.19 shows the density of states expected for semiconducting nanotubes. Experimentally, this



**Fig. 8.19.** Density of states and electronic band structure of a semiconducting nanotube near the Fermi level, plotted in reduced units  $E/E_0$ , where  $E_0 = 3a_0\beta/d_{\text{NT}}$ . From the webpage of C. Schönberger

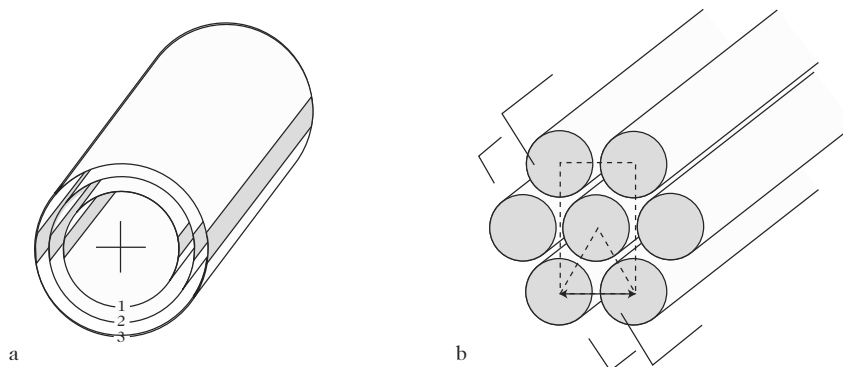


**Fig. 8.20.** Dependence of the energy gap  $E_g$  on the diameter of the nanotube [48]

dependence has been observed in measurements made by tunnel spectroscopy, as can be seen from Fig. 8.20. This is an indication, along with other spectroscopic studies carried out on metallic nanotubes which lead to estimates of the density of states, that the calculations of the electronic band structure of nanotubes made by the rather simplistic tight-binding method do in fact produce realistic approximations. However, it should be noted that, especially for nanotubes with small diameters, there are significant discrepancies between the electronic band structures calculated as above and the actual structures. These discrepancies result mainly from the curvature of the nanotube.

### 8.4.3 Self-Organisation of Nanotubes

Depending on the conditions under which the nanotubes are synthesised (see Sect. 8.4.5), they can self-organise themselves during the synthesis according to two possible self-assembly modes. In the first mode, shown schematically in Fig. 8.21a, the tubes are nested inside one another rather like Chinese boxes,



**Fig. 8.21.** The two modes of self-assembly for carbon nanotubes. (a) Structure of a multiwalled nanotube (MWNT). (b) Crystallised bundle of single-walled tubes

to form what are known as multiwalled nanotubes (MWNT) [3]. The number of walls and their diameter can vary greatly. In the second mode, illustrated in Fig. 8.21b, the nanotubes are single-walled (SWNT) with very similar diameters, in such a way that they may assemble to form bundles, sometimes called ropes [12]. In each bundle, the tubes pack together compactly to form a periodic arrangement with triangular symmetry. There may be as many as a few dozen in one bundle, with diameters in the range 0.6–1.5 nm, depending on the conditions of synthesis.

In the two types of self-assembly, the distance between neighbouring tubes is roughly equal to the distance between two sheets in graphite. This suggests that such an assembly of nanotubes does not modify the type of chemical bonds, these remaining the same as in graphite.

#### 8.4.4 Chemical Varieties of Nanotubes

Nanotubes can form from elements other than carbon. In fact, it may be possible to synthesise any element or chemical compound with layered (or lamellar) structure in a tubular form, provided that suitable conditions are established for the synthesis. For example, it is known that, under certain physicochemical conditions, the lamellar structure of minerals becomes unstable, whereupon they may transform into a close-packing of nanotubules. Likewise, the lamellar structure of certain liquid crystals can encounter instabilities which lead to the formation of nested vesicles, reminiscent of an onion.

Returning to nanotubes, the substance with structure closest to graphite is hexagonal boron nitride. To describe this structure, boron and nitrogen atoms alternate at the vertices of the hexagons in the graphene planes. It has in fact been possible to synthesise nanotubes of boron nitride by adapting methods used to synthesise carbon nanotubes [13, 14]. These nanotubes have the same structure as their carbon counterparts, except at their ends. Indeed,

the introduction of a pentagon into the hexagonal lattice of BN induces the formation of B–B or N–N bonds which are energetically unfavourable. This difficulty is overcome by introducing squares instead of pentagons. Each end of a nanotube is closed by introducing three squares. The most favourable configuration from the energy standpoint is obtained when the squares are distributed at the vertices of a triangle which defines a plane facet perpendicular to the tube axis [13]. Boron nitride nanotubes thus have extremely flattened ends, distinguishing them rather characteristically from their carbon counterparts. Finally, these BN nanotubes self-assemble into multiwalled structures during synthesis [13] or bundles of single-walled tubes [14].

The possibility of producing these nanotubes has opened the way to the synthesis of different types of heteroatomic nanotubes based on carbon, nitrogen and boron. Finally, nanotubes have also been synthesised from the lamellar compounds  $\text{WS}_2$  and  $\text{MoS}_2$  [15].

#### 8.4.5 Synthesis of Nanotubes

Carbon nanotubes may exist in the natural state, but at the present time only synthetic nanotubes have been observed. Since the original discovery by S. Iijima [3], various devices have been tested for this synthesis, with the twofold aim of producing new objects and of finding methods which will eventually make it possible to produce nanotubes on a large scale and in a controlled way. Two types of device can be distinguished here, differing principally in the temperatures required in each case.

##### High Temperature Synthesis

The first method involves high temperatures as a matter of principle, since it consists in vaporising graphitic carbon (graphite sublimes at  $3\,200^\circ\text{C}$ ) and then condensing it in a vessel where there is a strong temperature gradient and an inert gas such as helium or argon at a partial pressure that is typically around 600 mbar. The various methods using this basic idea can be distinguished from one another by the process used to vaporise the graphite.

In the Krätschmer–Huffmann process [2], used historically by S. Iijima [3], an electric arc (30–40 V, 100 A) is set up between two graphite electrodes. One electrode, the anode, is consumed to form a plasma at a temperature of up to  $6\,000^\circ\text{C}$ . This plasma condenses on the other electrode, the cathode, in a deposit containing the nanotubes [3]. Not only is the electric arc process very simple and cheap, but it is easy to carry out and modify to obtain different types of nanotube. One simply adapts the composition of the electrodes and the gas in the vessel. All the original syntheses of carbon and boron nitride nanotubes, either multiwalled or single-walled, were achieved in this way.

In the case of carbon, multiwalled nanotubes and bundles of single-walled nanotubes are obtained under radically different conditions. If the cathode is made from pure graphite, the nanotubes are exclusively multiwalled and form

directly in the vapour phase in the hottest region of the arc, at a temperature of at least  $3000^{\circ}\text{C}$ . The deposit thereby obtained is localised on the front face of the cathode at the point of impact of the arc.

In order to obtain single-walled nanotubes, a metallic catalyst must be used. This catalyst is introduced into the graphite powder to a level of a few percent [16]. It can be a transition metal such as Ni, Co, Pd or Pt, or a metal belong to the rare earth series, such as Y or La, or else it can be a mixture of these. Adding a rare earth to nickel, for example, considerably increases the production yield and facilitates the growth of long bundles of single-walled tubes which may contain up to a few dozen tubes [17]. These nanotubes form in a colder region of the arc, on the periphery of the cathode, where they build up a fringe with a rubbery and filamentary appearance, not unlike a very dense spider's web.

The synthesis of boron nitride nanotubes requires slightly more delicate conditions. They cannot be obtained from boron nitride electrodes owing to the insulating nature of this material. This obstacle has been overcome by using hafnium boride electrodes, a metal compound with high melting point which is known to decompose in the presence of nitrogen to give boron nitride [13]. The nanotubes synthesised in this way are either multiwalled or single-walled.

The second vaporisation process, originally developed by R.E. Smalley and coworkers at the University of Houston (Texas) [4,12], consists in bombarding a target by high energy laser radiation. The conditions of synthesis and the nature of the nanotubes thereby produced depends on whether the laser is pulsed (Nd:Yag-type laser) [12] or continuous wave (1–5 kW  $\text{CO}_2$  laser) [18]. Under the impact of a pulsed laser (the frequency of the pulses generally varies between 8 ns and 10 ms), the target undergoes ablation to form small clusters which can only recombine into a crystal structure if the reactor is placed in an oven heated to at least  $800^{\circ}\text{C}$ . These temperatures are not sufficient to produce multiwalled nanotubes and this method is therefore only used to synthesise bundles of single-walled tubes by ablation of a graphite target containing a metal catalyst based on Ni, Co and Fe.

When a CW laser is used, the target is heated by the laser to temperatures above 3000 K and gradually vaporises. The gas circulating in the vessel is overheated in the vicinity of the target surface and plays the role of a local oven. An external oven is not necessary in this case and the pressure and flow of carrier gas can be adjusted in order to control the temperature gradients. As in the case of the electric arc, MWNTs are obtained by vaporising pure graphite and SWNTs by vaporising graphite doped with a metal catalyst. This setup has a decisive advantage for the synthesis of insulating nanotubes, since it can be used to synthesise exclusively single-walled boron nitride nanotubes by using a boron nitride target or a pure boron target in a nitrogen environment [14].

Finally, a novel way of vaporising graphite consists in using solar energy. To do so, the solar radiation must be concentrated on a target in such a way

as to reach the vaporisation temperature of graphite. The feasibility of this process for synthesising both MWNTs and SWNTs has been demonstrated by D. Flamant and coworkers at the Odeïlho solar oven in cooperation with the University of Montpellier II (France) [19].

All these processes are able to produce in one experiment and on the laboratory scale anything between a few hundred milligrams and one gram of unpurified nanotubes. The ease of implementation of the electric arc method contributed greatly to the initial development of nanotube research. But this flexibility has to be balanced against the complexity of the processes occurring during the synthesis, which make them difficult to control and study *in situ* [20]. In contrast, the laser processes are costly but involve a relatively limited number of control parameters, making it possible to study and model the conditions of synthesis. Several studies have been carried out for pulsed laser [21] and CW laser [22] systems. The operating mode of the solar oven is very similar to vaporisation by CW laser and is similarly well suited to *in situ* studies [19].

Finally, although it is difficult to envisage large scale production of nanotubes using these high temperature methods, production units on scales beyond the laboratory scale are currently under investigation: in Odeïlho via the solar approach, and in Montpellier since 2001 via the electric arc approach.<sup>8</sup>

### Moderate Temperature Synthesis

The second approach to nanotube synthesis operates at moderate temperatures. It is an adaptation of catalytic or pyrolytic methods (CCVD) traditionally used to synthesise carbon fibres, as mentioned earlier [7]. The idea behind CCVD methods is to decompose a carbon-bearing gas at the surface of particles of a metal catalyst in an oven heated to between 500 and 1100°C, depending on the gas used. The carbon released by the decomposition of the gas then precipitates onto the surface of the particle and this condensation results in the growth of graphitic tubular structures. The carbon-bearing gas can be carbon monoxide (CO) or a hydrocarbon such as acetylene, methane, etc. The metal catalyst is a transition metal like iron, nickel, or cobalt. One delicate aspect of these techniques is the preparation and control of the size of the catalyst particles. They must be of the order of a few nanometers for the synthesis of nanotubes. In order to obtain single-walled nanotubes, the particles are obtained by reducing organometallic compounds such as ferrocene [23] or specific metal oxides [24]. They are then deposited on a ceramic supporting material (silica or aluminium oxide) [24, 25] or blown into the reaction chamber where the reaction with the carbon-bearing gas takes place [23].

Depending on the operating conditions (oven temperature, pressure and flow rate of the gases, size of particles), the synthesis results in either MWNTs or SWNTs. Generally, despite the diversity of the parameters, SWNTs are

<sup>8</sup> See the webpage [www.nanoledge.com](http://www.nanoledge.com) of the company Nanoledge.



synthesised at higher temperatures than MWNTs and using smaller catalyst particles. The MWNTs obtained by these methods often exhibit a distinctly inferior quality of graphitisation than those obtained by high temperature techniques. On the other hand, they have much more uniform geometric characteristics, i.e., length and diameter, and this is an advantage. In addition, tube growth can be oriented and localised by synthesising the tubes on dots of catalyst arranged with some definite geometry on a substrate [26, 27]. This possibility provides very interesting prospects for certain applications, as we shall see later. Moreover, medium-temperature processes can be scaled up to the production levels of carbon fibres, something that is much more difficult to achieve using high temperature channels. To this end, the University of Houston (Texas) has developed a device using carbon monoxide which currently produces 10 g of unpurified SWNTs per day and which is in the process of being industrialised by Carbon Nanotechnologies Inc. (CNI).<sup>9</sup>

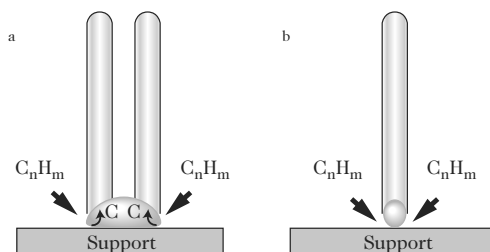
#### 8.4.6 Growth Mechanism for Carbon Nanotubes

The two self-assembly modes for nanotubes are mutually exclusive and occur under radically different conditions of synthesis. In the high temperature channels, bundles of single-walled tubes can only be obtained by mixing a few percent of a metal catalyst into the graphite powder. This catalyst is a transition metal such as Ni, Co, Pd, or Pt, or a metal belonging to the rare earth series, such as Y or La, or a mixture of these metals. In the medium-temperature approaches, the type of nanotube assembly seems to be controlled by the temperature and the size of the catalyst particles [23–25]. Numerical simulations using *ab initio* molecular dynamics (MD) methods allow one to analyse the growth mechanisms on the microscopic scale and to understand why the growth of single-walled nanotubes requires different conditions of synthesis to the growth of multiwalled nanotubes [28].

In a highly schematic way, one may say that an open SWNT will close spontaneously by establishing bonds between atoms located on the rim of the nanotube. This evolution is too fast to allow incorporation of new carbon atoms into the nanotube lattice, whereupon this instability makes it impossible for an isolated tube to grow in the vapour phase. However, such growth is possible for MWNTs, because dynamic bonds form between the rims of adjacent tubes, thereby stabilising the open tubes and allowing the incorporation of new carbon atoms from the vapour. This analysis allows one to understand why it is possible to synthesise MWNTs in the vapour phase at very high temperatures. For SWNTs, *ab initio* calculations show that the presence of metal atoms such as cobalt at the end of the tube tend to stabilise it, allowing once again the incorporation of further carbon atoms.

This analysis does not provide a mechanism for the formation and growth of SWNTs, but it does provide some explanation as to why the catalyst is

<sup>9</sup> This is the HipCo process, for which CNI has the exclusive licence. The webpage is [carbonnanotech.com](http://carbonnanotech.com).



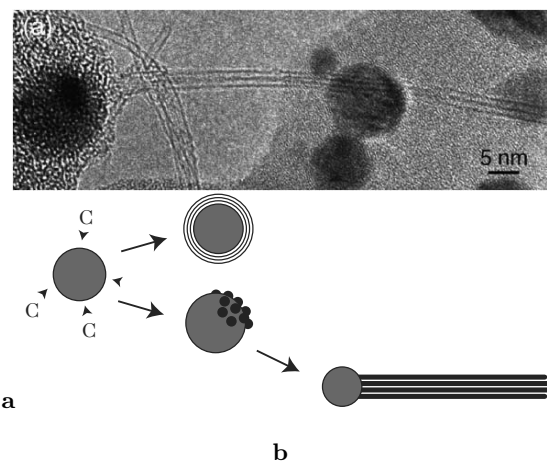
**Fig. 8.22.** (a) Formation and growth of carbon whiskers by CVD techniques [6,7,30]. (b) Adaptation to the formation of an SWNT using CVD techniques [25]

essential here. Detailed electron microscope analysis of the morphology of SWNT ropes obtained using different methods of synthesis have recently allowed a more precise assessment of the role played by the catalyst [29]. It has been established that, in all cases, the nanotubes form from small catalyst particles and that two situations must be distinguished, as illustrated in Fig. 8.22, depending on whether the tubes have grown parallel or perpendicularly to the particle surface. In the first case, the diameter of the nanotube is conditioned by the diameter of the particle, which is generally encapsulated at the end of the tube, whilst in the second case, there is no such correlation, since several tubes may grow from the same particle.

The first case, observed for tubes synthesised by CVD, corresponds to the well known situation of carbon whiskers, and the growth mechanisms shown schematically in Fig. 8.22 must apply considering the SWNT as the smallest possible filament. These mechanisms proceed by chemisorption and decomposition of the carbon-bearing gas at the surface of the catalyst, diffusion of the released carbon in the particle, and segregation and graphitisation parallel to the particle surface [6, 7, 30]. The carbon concentration gradient between the surface and the bulk of the particle ensures continuity of the process for as long as the carbon supply lasts and the catalyst remains effective. The catalyst is thus an element which favours chemisorption of the carbon-bearing gas and which segregates the carbon in the solid state so as to allow formation of graphitic structures at low temperatures. Metals like Fe, Ni, or Co fulfill these criteria.

This mechanism, which explains the correlation between the diameter of the tube and the diameter of the particle, does not apply to perpendicular growth observed mainly in high temperature techniques. In this case, electron microscope observations strongly suggest a formation mechanism of type vapour–liquid–solid [31, 32], illustrated in Fig. 8.23. This phenomenological model, supported by *ab initio* calculations and consistent with *in situ* studies of the vapour phase made for the laser and solar vaporisation techniques [19–22], rests upon the thermodynamic properties of systems such as Ni–C and Co–C.

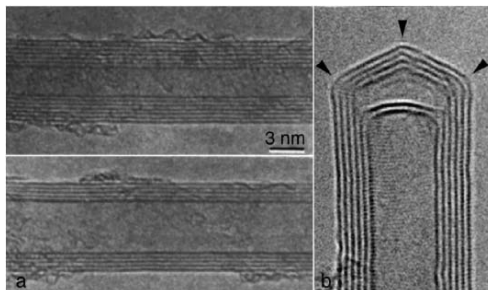
According to this model, nanotube formation follows a specific kinetic pathway wherein carbon is segregated at the surface of small liquid particles



**Fig. 8.23.** (a) Transmission electron microscope image of a nanotube bundle obtained by a CW laser reactor [29]. The bundle emanates from a catalyst particle. All the nanotubes emerge from the same particle and growth has occurred perpendicularly to the surface of this particle. (b) Formation and growth of carbon nanotubes according to the vapour-liquid-solid model [32]

of catalyst (10–20 nm in size) that are saturated with carbon. The first step is condensation of the carbon and metal vapours. Given the difference in vapourisation temperatures of carbon and metals such as Ni or Co, it is the carbon that condenses first to form low density clusters, followed by the metal, which forms small liquid metal particles with sizes that can be controlled by local cooling conditions. In the liquid state, these particles can dissolve large quantities of carbon. During cooling, the solubility threshold falls off to become practically zero at the solidification point (around 1 200°C). The supersaturated carbon segregates out by diffusing to the surface of the particle and forming a graphitic structure. Depending on the local segregation conditions, it can either form continuous layers of graphite encapsulating the particle in a kind of shell, or it can form islands which become the seeds for nanotubes. This nucleation hypothesis is strongly supported by direct observations of these seeds at the surface of the particles [32]. It implies that the formation of nanotubes is correlated with the solidification temperature of the metal, which does seem to be borne out experimentally. Finally, this mode of formation involves nanotube growth from one end. According to this assumption, the length of the tubes should be directly related to the stability of local conditions, in particular, the temperature and carbon supply, and it would explain the need for an oven in the pulsed laser setup.

Although this mechanism seems at present to be fairly well established from a qualitative standpoint, much remains to be done in this field before we are able to conduct the synthesis of a tube of given configuration.



**Fig. 8.24.** TEM observation of a multiwalled nanotube. (a) Micrographs of two nanotubes containing 5 and 7 walls. Each wall is visualised by two black lines arranged symmetrically on either side of the nanotube axis. Adapted from [1]. (b) Image of the end of a nanotube. *Arrows* show the positions of pentagons which allow the walls to close [3]

#### 8.4.7 Observation of Nanotubes

The main instrument for observing and identifying nanotube structures is the transmission electron microscope (TEM). As we saw in the introduction, it was this device that first allowed the discovery of nanotubes. There are many different interactions between electrons and matter, and each can provide a different type of information via the electron microscope. The imaging and diffraction modes provide information concerning morphology and microstructure from the micron scale to the nanoscale, as well as information about atomic arrangements (crystal structure, defects), whilst the spectroscopic modes (EDX, EELS) allow one to identify the chemical nature of the atoms and their spatial distribution within the sample (elemental profiles and chemical mapping). These different modes have been used together in a complementary way to study the structure and chemistry of nanotubes.<sup>10</sup> Two examples are discussed here by way of illustration.

The structural and chemical analysis of multiwalled nanotubes is illustrated in Fig. 8.24. The image in Fig. 8.24a shows a nanotube projected perpendicularly to its axis. A tube projected in this way is visualised by lines parallel to and symmetrically arranged on either side of the tube axis, these being the projection of the parts of the tube tangential to the bundle, shaded in the diagram of Fig. 8.21a. The image thus shows that the object is made up of equidistant concentric layers and it immediately reveals their number and spacing. However, it does not allow us to conclude that the walls have axial symmetry. The cylindrical geometry of the layers has been checked by various methods, in particular by chemical analyses using EELS, scanning a subnanometric electronic probe along a cross-section of a nanotube. This

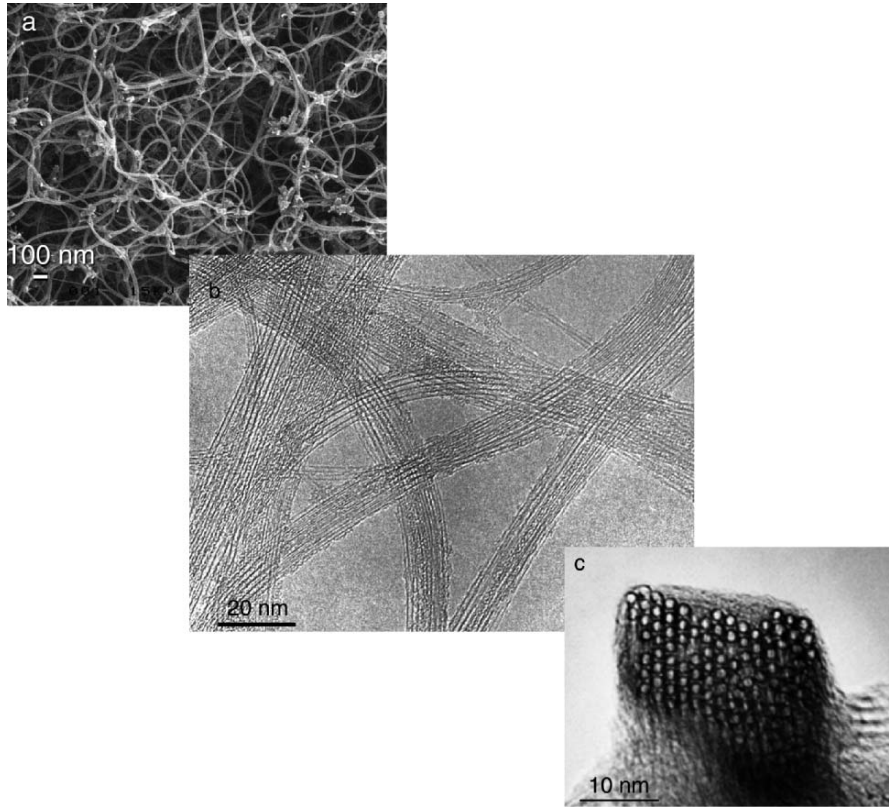
<sup>10</sup> For a review of the use of diffraction and imaging in the study of nanotubes, see the very detailed report by S. Amelinckx et al. [33]. These aspects and the use of energy loss spectroscopy are also discussed in detail in [54].

yields a chemical profile which is an image of the atomic density of the object projected perpendicularly to its axis. This profile has a butterfly shape typical of a tubular structure, the two maxima being due to the part of the tubes tangential to the electron beam and the minimum occurring in line with the axis due to the part of the tubes perpendicular to the beam. Figure 8.24 shows an example of the end of a multiwalled tube. The black lines imaging the walls show variations in the curvature which reveal the presence of pentagons required to close the walls at the tube end.

Figure 8.25 examines the structure of bundles or ropes of single-walled tubes. Unlike the MWNTs which are straight, SWNT ropes are flexible and bend very easily to form a tangled mass which looks at low magnifications like a plate of spaghetti or a tangle of hair (see Fig. 8.25a). The fact that these ropes are able to bend means that they can be observed either in perpendicular projection (Fig. 8.25b) or in parallel projection (Fig. 8.25c) with respect to their axis. The latter projection is particularly interesting, because it provides a direct image of the cross-section of the bundle and the tubes making it up, revealing in this way their intimate organisation. In Fig. 8.25c, each black circle is the image of a single-walled tube. One can immediately see the homogeneity of the diameters, the periodicity of the packing, and the number of tubes in the bundle. When the tubes are viewed perpendicularly to their axis, the image shows a set of equidistant fringes corresponding to the projected rows of tubes. This mode allows a detailed study of the architecture of the tangled ropes.

Although the observations in Figs. 8.24 and 8.25 reveal the tubular structure of the nanotubes and the way they assemble, they give no clues as to their helicity. To investigate this question, the positions of the carbon atoms in each tube must be imaged directly, a task that encounters two difficulties. Firstly, the microscope must have resolution better than 0.2 nm, and secondly the projected view makes it difficult to establish the positions of atoms in a stack of tubes. The solution is to obtain the diffraction pattern in a plane parallel to the tube axes, which can be achieved using electron diffraction [33]. Not only is it possible to determine the helicity distribution of a set of tubes in this way, but in certain specific situations, such as the observation of double-walled tubes, it is even possible to determine the indices  $(n, m)$  of the various tubes [34]. It can be concluded from a careful examination of a large number of tubes that tubes obtained by high temperature synthesis have no particular helicity, whatever assembly mode they adopt, whereas CVD tubes may exhibit a uniquely defined helicity within a given bundle [35].

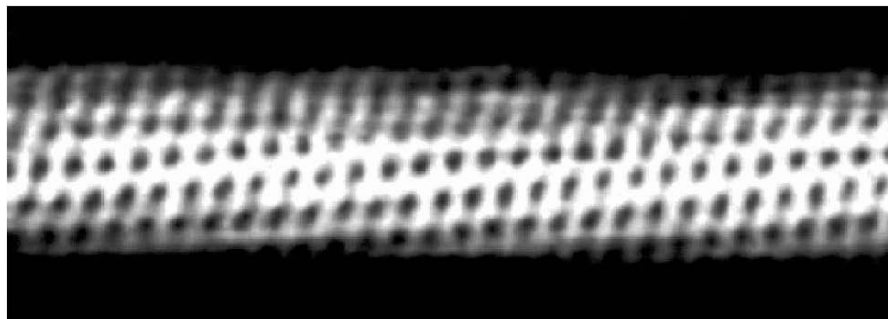
An alternative and elegant solution to the helicity problem consists in imaging the surface of the nanotubes on the atomic scale using a scanning tunneling microscope (STM). The idea of this microscope, developed in 1987 by G. Binnig and H. Rohrer, is to use a very fine metal tip as a probe. A bias potential of a few volts is applied between the sample surface and the tip, and the latter is brought to within a few tenths of a nanometer from the surface in order to scan across it. When the tip passes near a surface atom, electrons



**Fig. 8.25.** Observation of SWNT ropes. (a) Scanning electron microscope image of SWNT ropes. Courtesy of L. Vaccarini, University of Montpellier, France. (b) TEM micrograph of ropes observed perpendicularly to their axis. *Black lines* correspond to projected rows of tubes. Courtesy of J. Gavillet, Onera-CNRS, France. (c) TEM micrograph of a rope looking along its axis. Each *black circle* corresponds to one nanotube. Adapted from [4]

can cross between this atom and an atom in the tip, overcoming the potential barrier by the quantum tunneling effect and setting up the so-called tunnel current. The current mapping obtained by scanning the tip across the sample produces an image of the surface atoms. An image of a nanotube obtained by C. Dekker and coworkers at the University of Delft using this technique is shown in Fig. 8.26 [36, 37]. The intensity maxima correspond to the atoms and their distribution reveals the hexagonal carbon lattice. This lattice has no particular orientation with respect to the tube axis, thereby indicating its chirality in complete agreement with electron diffraction analyses.

Finally, note that Raman spectroscopy can provide a global analysis of the nanotube structure which perfectly complements local electron microscope analysis. Schematically, the tubes exhibit low frequency radial breathing



**Fig. 8.26.** STM image of a chiral nanotube showing the hexagonal arrangement of the carbon atoms. Adapted from [27]

modes which provide the diameter distribution of the tube samples, and medium frequency modes sensitive to the tube helicities, with a resonance that depends on the laser wavelength used. This kind of spectral analysis is used routinely to assess the quality of samples and compare one sample with another.

#### 8.4.8 Properties of Nanotubes

The properties of nanotubes are a direct result of their structural inheritance from graphite. The planar and directed chemical bonding in graphite makes it a chemically very stable and highly anisotropic solid. Most of its properties arise from its plane honeycomb structure which is precisely what provides the basic skeleton of the nanotube [5]. Properties specific to the nanotube stem from the perturbations to this plane due to curvature and the reduction in dimensionality. The properties of the nanotube thus arise by adapting the properties of graphite to the conditions imposed by rolling up the graphene sheets. They may be classified into electrical conduction properties, mechanical properties, thermal properties, and chemical properties.<sup>11</sup>

The electrical properties of nanotubes follow from their electronic band structure. As we saw earlier, nanotubes can be semiconducting or metallic depending on their structure and in particular, the roll-up vector. Measurements carried out on nanotubes do in fact reveal a variety of resistive behaviour which confirms the theoretical conclusions drawn previously, predicting the existence of semiconducting and metallic nanotubes. In particular, at low temperatures, single-walled metallic nanotubes, either individually or in small bundles, behave as quantum wires, i.e., conduction occurs via well separated, discrete electronic levels that are quantum-mechanically coherent over lengths of a few hundred nanometers. At low enough temperatures, such wires constitute an extended quantum dot.

<sup>11</sup> The references [5, 51–53] deal in some detail with the properties of nanotubes.

More precisely, in the case of metallic nanotubes that are weakly coupled to electrodes at a temperature of a few kelvins, the conductance at low voltages is suppressed below a few millivolts. Moreover, peaks appear clearly when the variation of the conductance is plotted against the electrostatic potential applied to an electrode near the nanotube. These results have been interpreted in terms of the one-electron charge transfer and resonant transport via the quantised energy levels of the nanotubes. Apart from the quantisation arising through the structure of the nanotube itself, its limited lengthwise spatial extension imposes a further quantisation condition corresponding to observed confinement effects. As a matter of fact, the structural simplicity and chemical stability of metallic nanotubes makes them a model form of molecular quantum wire.<sup>12</sup>

It should be stressed that an excellent correlation has been established between the electronic and structural properties of carbon nanotubes using near-field microscopy, especially STM [53].

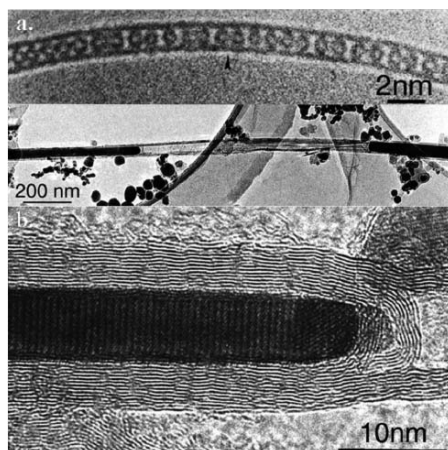
Let us return briefly to the case of metallic single-walled nanotubes in order to give some orders of magnitude as far as their conductance is concerned. If no precautions are taken, a resistance measurement made (in two-point configuration) on a nanotube connected between two metal electrodes gives a value of the order of 1 M $\Omega$ , which is essentially fixed by the contacts. It has been shown [45] that ohmic contacts can be made by suitable evaporation of a metal on the nanotube, in such a way that one may approach the theoretical value  $h/4e^2$  of the conductance, i.e., 6.5 k $\Omega$ , due to the existence of two subbands at the Fermi level. The possibility of establishing good electrical contacts has also led to the demonstration that multiwalled nanotubes can become superconducting at very low temperatures and in certain conditions [46].

Naturally, the semiconducting properties of nanotubes have been used to develop electronic components presented below and discussed further in Chap. 13. We end this discussion of electrical properties by mentioning that nanotubes can carry quite remarkable current densities, of the order of  $10^{10}$  A/cm<sup>2</sup>, at least two orders of magnitude greater than metals.

The second class of properties we mentioned above were mechanical properties. Due to its structural anisotropy, graphite has a very high elastic modulus (measuring its resistance to deformation) in the hexagonal plane, reaching values of  $10^{12}$  Pa or 1 TPa, but much lower values out of the plane, at  $4 \times 10^9$  Pa. The nanotube inherits the mechanical capacities of graphene, and even enhances them, since elastic moduli of 1 TPa have been calculated and measured [38]. It combines this resistance to deformation with a high level of flexibility, due to the partial  $sp^2$ – $sp^3$  hybridisation of the C–C bond. Various experiments have shown that the nanotube has an incredible ability to bend to considerable angles, and also to deform and twist about its axis [38].

<sup>12</sup> The reader is referred to Chaps. 9 and 13 on nanowires and molecular electronics, respectively, and also to [52–55] for a more extensive bibliography.





**Fig. 8.27.** TEM micrographs of filled tubes. (a) Single-walled nanotube filled with  $C_{60}$  molecules. Adapted from [39]. The tube has been opened by oxidation and exposed to a gas of  $C_{60}$  molecules at  $400^{\circ}\text{C}$ . The molecules enter the tubes by capillary forces. (b) Multiwalled tube filled with chromium sulfide crystals, magnified in (c) (Photos A. Loiseau.) The tube and filling were obtained by direct synthesis using the electric arc method with an anode made from C, S, and Cr [40]

However, these elastic properties, which have indeed been confirmed experimentally for individual SWNTs, are considerably downgraded in MWNTs, and all the more so as the number of walls increases, as well as in SWNT ropes.

Finally, nanotubes also have very attractive chemical properties. As shown in Fig. 8.27, they can be filled by capillary effect with fullerene molecules [39], or with crystal compounds to produce encapsulated nanowires [40]. These compounds can be metals, sulfides or metallic chlorides. It is also possible to tether molecules to the nanotube surface in order to functionalise it and use it as a support for synthesis. Protein crystals have also been produced. Finally, nanotubes can even be intercalated and doped electronically like graphite by inserting electron donor or acceptor elements between the walls of MWNTs or in the channels between the tubes in a bundle. This intercalation serves to control the Fermi level of the nanotubes and modulate their electronic properties.

#### 8.4.9 From Science to Applications

Going beyond fundamental research on nanotube physics and chemistry, we may ask what the future holds in store as far as applications are concerned. When we ask how their properties may be exploited, we may distinguish those applications using the nanotube as a nano-object and those using it

as a component of a composite material whose functionalities are thereby reinforced or modified.<sup>13</sup>

### Electronics

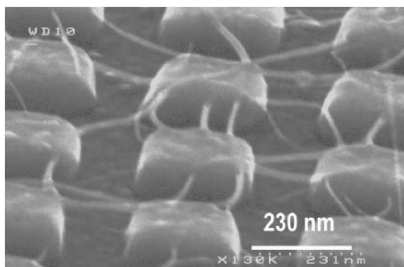
At the top of the list, in the field of electronics, nanotubes may be considered as a model for conducting wires or functional components in the years 2010–2015. Indeed, the trend towards miniaturisation of electronic devices<sup>14</sup> will require nanoscale components within 10 to 15 years from now. Although such components will certainly be made from silicon and standard semiconductors, certain physical or technological limitations require the development of complementary solutions. One possibility is the emergence of molecular electronics, the subject of Chap. 13. It would be risky to claim today that the nanotube will play a part in such electronics, or to try to guess at what level it might be involved. We may be fairly certain that the materials to be used in future devices have yet to be discovered. However, the nanotube is a molecular object that is already available as a training ground for working with nanoscale components, with which we may gain practice in manipulating them, studying their physical and chemical behaviour, and designing functional devices.

The nanotube thus serves as an ideal model tool in this apprenticeship, because it is chemically simple and stable, and it is endowed with astonishing electronic properties. It has allowed very rapid progress to be made in this field. Within barely 5 years, the various functional and logical components of integrated circuits such as the field-effect transistor, the Schottky barrier, the Schottky diode, the  $p$ - $n$  junction, not to mention memory cells, have already been fabricated using semiconducting nanotubes. What is more, they perform as well as or even better than their standard semiconducting counterparts for comparable geometries. In addition, conducting nanotubes could provide an alternative to today's metal wires, which will be difficult to miniaturise beyond a certain point. By virtue of the chemical stability of its C–C bond and its thermal properties, a nanotube can sustain currents that would melt a copper wire of the same dimensions. The idea of replacing copper interconnects in today's integrated circuits with interconnects based on carbon nanotubes is currently being investigated.

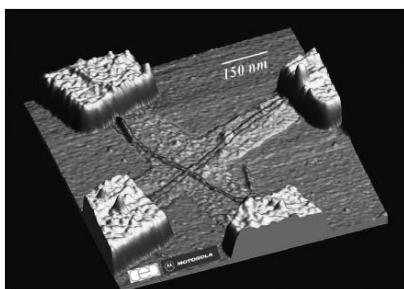
However, we are still a long way from understanding the physical behaviour of these devices, one problem being to establish the role played by the contacts between nanotubes and circuit elements such as electrodes which have dimensions so much bigger than the nanotubes. Quite generally, the extent to

<sup>13</sup> The reader is referred to [41] for an account of their properties and corresponding prospects for applications.

<sup>14</sup> This process of miniaturisation concerns in particular all portable devices, such as portable phones and pocket computers, but also memories for next-generation computers. Miniaturisation not only gains space, but also increases transmission efficiency and energy saving.



**Fig. 8.28.** Scanning electron microscope image of suspended nanotubes synthesised by hot filament CVD on Si dots coated with Co particles. Photo courtesy of LEPES+CRTBT [43]

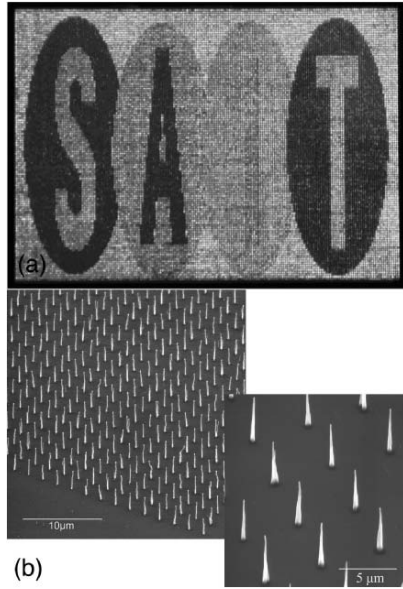


**Fig. 8.29.** AFM image of crossed nanotubes. The nanotubes are self-assembled on tracks which serve to localise the deposit. The tracks ( $150\text{ nm} \times 0.7\text{ nm}$ ), *grey* in the figure, consist of monomolecular layers of silane molecules terminated by amine functions, formed through a mask made by electron beam lithography. Photo courtesy of E. Valentin, S. Auvray LEM CEA-Motorola [47]

which nanotubes can be exploited in future electronic devices will depend in part on our ability to control the helicity of the tubes during synthesis and our ability to integrate nanotubes into devices and circuits, i.e., to develop two- and three-dimensional architectures involving these entities.

The experiments discussed above are extremely sophisticated, using the most advanced laboratory equipment and techniques, such as nanolithography, atomic force and scanning tunneling microscopy, to mention but a few. There is a long way to go from this experimental level of fabrication to mass production in millions of components! Concerning the question of integration, one approach that appears very promising at the present time involves using the CVD technique to synthesise nanotubes locally on Si dots in a circuit, on which catalyst particles have been deposited as shown in Fig. 8.28 [27,43]. Another approach consists in implementing nanotube self-assembly techniques driven by chemical recognition (see Fig. 8.29) [47]. The various points concerning the use of nanotubes for electronics are discussed in more detail in Chap. 13.

The problem of controlling the helicity, which determines the conducting or semiconducting nature of the tubes, remains formidable. At the present time, synthesis methods are totally unselective and our current understanding of the formation mechanisms does not yet provide any information that could help us to control this aspect during synthesis. Current efforts are based rather on chemical manipulation of the tubes, which means functionalising them selectively depending on their conduction behaviour and then separating the different phases.



**Fig. 8.30.** (a) Photograph of a prototype flat screen using multiwalled nanotubes, with  $576 \times 242$  pixels, made by Samsung. (b) Scanning electron microscope images of an array of oriented multiwalled nanotubes, synthesised by a plasma CVD technique using as a starting point an array of Ni catalyst dots. *Insert:* Magnified view showing that each dot gives rise to a single nanotube. The distance between the dots has been determined to avoid screening of the emission current arising from the different tubes. Photo courtesy of Thales

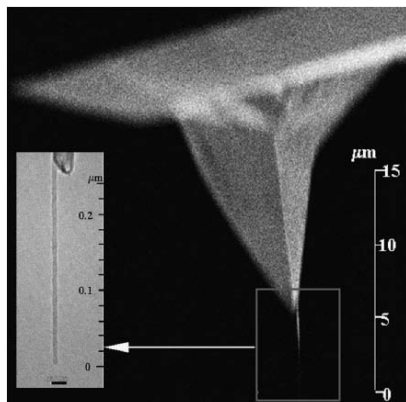
### Electron Emission in an Electric Field

A nanotube emits electrons through its tip by the tunnel effect<sup>15</sup> when placed in an electric field with field lines parallel to the nanotube axis [44]. Owing to their aspect ratio, thermal stability and mechanical qualities, nanotubes are extremely effective emitters: aspect ratios greater than 1 000, an extremely low emission threshold bias, the ability to supply large currents that remain stable over long periods. It is no surprise therefore that several applications of this feature have already reached the industrialisation stage for cold cathode devices and flat television screens (see Fig. 8.30a) and that these applications are the ones attracting most interest from the industrial sector at the present time. Here again, CVD synthesis techniques offer decisive advantages because it is possible to synthesise custom-built oriented arrays of tubes starting from an array of catalyst dots obtained by lithographic methods, as illustrated in Fig. 8.30b.

### Probes for Nanoscale Systems

The ends of nanotubes are extremely finely tapered. Combined with their electrical and mechanical properties, this makes them ideal tips for probing

<sup>15</sup> Electrons in levels close to the Fermi level are extracted by the tunnel effect. This extraction occurs beyond a threshold voltage bias applied between the emitter and a counterelectrode. The value of the bias depends on the electronic structure of the material. The emitted current grows as the aspect ratio increases, according to the classical Fowler–Nordheim law [45].



**Fig. 8.31.** Tip of an atomic force microscope with a multiwalled nanotube stuck onto it. *Insert:* TEM image of the nanotube. Adapted from [4]

the physical and physicochemical properties of systems with nanoscale dimensions, such as nanopatterned surfaces or molecules, or for studying chemical reactions. They are already used in near-field microscopy. These tips considerably enhance the resolving power of such instruments compared with standard tips, because their very small radius of curvature allows one to overcome the problem of convolution of the tip with the imaged object. As can be seen from Fig. 8.31, these tips are obtained by fixing a multiwalled nanotube at the end of a standard tip. The nanotube is fixed either by sticking it on or by growing it directly on the standard tip by CCVD. Such tips are already commercially available.

### Nano-Objects for Detection, Energy Storage, Chemical Synthesis

As mentioned above, the geometric characteristics of nanotubes (surface, interior cavity, interwall space) make them ideal candidates for chemical manipulation such as gas adsorption, tethering of molecules or clusters, confinement of molecules within the cavity, and intercalation. Some of these manipulations can be carried out with other carbon-bearing structures, such as graphite for intercalation, or carbon black for adsorption, whilst others are specific to the tubular geometry.

Applications based on the chemical use of these nano-objects concern hydrogen storage, energy storage, chemical sensors and actuators, and possibilities for functionalisation chemistry. Despite the widely publicised announcement of revolutionary possibilities a few years ago, investigations into the storage of hydrogen for fuel cells now seem to converge upon a very limited value of 1–3%, below the threshold fixed by the car industry in particular. As far as energy storage is concerned, the most promising system seems to be the supercapacitor, which is in fact a composite system with a conducting polymer. Finally, nanotube chemistry, i.e., the possibility of tethering molecules to the surface, functionalising, filling, and so on, is still in its early stages.

However, it represents an extremely important research direction in the short and mid-term, because it represents a training ground for the general problem of controlled manipulation of nanotubes which will be vital for many applications, e.g., sorting helicities for molecular electronics.

### Materials for Every Purpose

The word ‘material’ here refers to bulk materials, with macroscopic dimensions, in which nanotubes have been incorporated. These are therefore composite materials, developed for a very wide range of applications depending on the property or properties of nanotubes that one wishes to exploit: structural materials in which the nanotube serves as a mechanical strengthener, electrically conducting materials, thermal materials, absorbent materials, or optically limiting materials for use in the bulk or as a surface coating. A good example of the potential of such materials has been developed by the automotive industry. Nanotubes are added to paints to make them slightly conducting and hence easier to apply to composite surfaces without the need for polar solvents, which are so bad for the environment.

In a general way, the fabrication of the material, the arrangement of the nanotubes within the material, and the nature of the nanotube–matrix interfaces are key problems revealed by studies carried out so far on conducting and structural composites. They will only be solved by a long-term effort. Once again, chemical manipulation for preparation and chemical functionalisation to adapt the nanotube to its environment should play a decisive role.

## 8.5 Conclusion

This review of applications is obviously not exhaustive, but it should be enough to show that nanotubes have a truly diverse potential for applications. The most remarkable attribute of these developments is that they often lie at the crossroads of several different disciplines. Better still, as our understanding of these objects moves forward, they continue to amaze us, with new experiments revealing yet more aspects that theory has failed to predict. This is a remarkable point because it is quite often the opposite situation which prevails after a discovery. Nanotubes are clearly more than just a fashionable subject in the wake of the nanoscience bandwagon. Indeed, they represent a very important subject of research which concerns extremely varied fields and problems in nanoscientific research and applications. This is because the nanotube unites an unparalleled assemblage of properties. It combines three structural characteristics – chemical simplicity (it is just carbon!), chemical and thermal stability (graphite chemical bond), and molecular dimensions – which mean that it can be simultaneously an excellent conductor or a semiconductor characterised by a gap equivalent to that of Si or Ge, as well as exhibiting considerable mechanical strength.

**Table 8.2.** Structural and electronic characteristics of carbon nanotubes

Characteristic	Graphene	General case $(n, m)$
Basis vectors of graphene lattice	$\mathbf{a}_1 = -\frac{\sqrt{3}}{2} a_0 \mathbf{i} + \frac{3}{2} a_0 \mathbf{j}$ $\mathbf{a}_2 = \frac{\sqrt{3}}{2} a_0 \mathbf{i} + \frac{3}{2} a_0 \mathbf{j}$ where $a_0$ is the length C-C	
Chiral vector $\mathbf{C}$		$\mathbf{C} = n\mathbf{a}_1 + m\mathbf{a}_2$ $\mathbf{T} = \frac{(2m+n)\mathbf{a}_1 - (2n+m)\mathbf{a}_2}{d_r}$
Vector $\mathbf{T}$ defining unit cell		$d_r = \begin{cases} \text{GCD}(m, n) & \text{if } (m-n) \text{ not multiple of } 3 \times \text{GCD}(m, n), \\ 3 \times \text{GCD}(m, n) & \text{if } (m-n) \text{ multiple of } 3 \times \text{GCD}(m, n), \end{cases}$
Diameter		$d_{\text{NT}} = \frac{1}{\pi} \ \mathbf{C}\  = \frac{\sqrt{3}}{\pi} a_0 \sqrt{(m^2 + nm + n^2)}$
Chiral angle		$\theta = \arctan \frac{\sqrt{3}m}{m+2n}$
Electronic structure	Semi-metal	Metallic if $n-m$ multiple of 3, semiconducting otherwise
Dispersion relation $e_i(k_x, k_y)$	$\pm \beta \sqrt{1 + 4 \cos\left(\frac{\sqrt{3}a}{2} k_y\right) \cos\left(\frac{a}{2} k_x\right) + 4 \cos^2\left(\frac{a}{2} k_x\right)}$ where $a = a_0\sqrt{3}$	$\pm \beta \sqrt{1 + 4 \cos\left(\frac{\sqrt{3}a}{2} k_y\right) \cos\left(\frac{a}{2} k_x\right) + 4 \cos^2\left(\frac{a}{2} k_x\right)}$ where $\mathbf{C} \cdot \mathbf{k} = 2\pi q$ and $q \in \mathbb{Z}$

**Table 8.2. (cont.)** Structural and electronic characteristics of carbon nanotubes

Characteristic	Armchair $(n, n)$	Zigzag $(n, 0)$
Chiral vector $\mathbf{C}$	$\mathbf{C} = n(\mathbf{a}_1 + \mathbf{a}_2)$	$\mathbf{C} = n\mathbf{a}_1$
Vector $\mathbf{T}$ defining unit cell	$\mathbf{T} = \mathbf{a}_2 - \mathbf{a}_1$	$\mathbf{T} = \mathbf{a}_1 - 2\mathbf{a}_2$
Diameter	$d_{\text{NT}} = \frac{3}{\pi} a_0 n$	$d_{\text{NT}} = \frac{\sqrt{3}}{\pi} a_0 n$
Chiral angle	$\theta = 30^\circ$	$\theta = 0^\circ$
Electronic structure	Metallic	Metallic if $n$ multiple of 3, semiconducting otherwise
Dispersion relation $\epsilon(k_x, k_y)$	$\pm \beta \sqrt{1 + 4 \cos\left(\frac{q\pi}{n}\right) \cos\left(\frac{a}{2} k_x\right) + 4 \cos^2\left(\frac{a}{2} k_x\right)}$	$\pm \beta \sqrt{1 + 4 \cos\left(\frac{q\pi}{n}\right) \cos\left(\frac{q\pi}{n}\right) + 4 \cos^2\left(\frac{q\pi}{n}\right)}$



Great changes are currently underway in the development of nanotube research. The first stage of acquiring a basic understanding has almost come to a close. Little by little, it is giving way to a phase of much more fundamental research into complex phenomena such as transport and emission, together with a concerted effort to develop applications. It should be noted, however, that applications in the pipeline today often require the fulfillment of a certain number of conditions which may or may not be difficult to achieve. Generally speaking, these conditions include controlling the synthesis of given tube configurations with production capacities adapted to the application; controlling purity and crystal quality; and developing suitable means of preparation for use in specific devices and for functionalising materials.

### Further Reading

The reader is referred to the general works [50, 51] for the fullerenes and carbon, and [52–56] for nanotubes. Reference [52] is a basic book on the physical properties of nanotubes, while [54] (still in press) is based on a CNRS summer school held in France in 2003.

### References

1. Kroto, H.W., Heath, J.R., O'Brien, S.C., Curl, R.F., and Smalley, R.: *Nature* **318**, 162 (1985)
2. Krätschmer, W., Lamb, L.D., Fostiropoulos, K., and Huffman, D.R.: *Nature* **347**, 354 (1990)
3. Iijima, S.: *Nature* **354**, 56 (1991)
4. <http://cnst.rice.edu/pics.html>: website for R. Smalley at Rice University (Houston, USA)
5. Bernier, P., and Lefrant, S.: *Le carbone dans tous ses états*, Gordon & Breach Science (1997)
6. Baker, T.K.: *Carbon* **27**, 315 (1989); Bacon, K.: *Appl. Phys. Lett.* **31**, 283 (1996)
7. Oberlin, A., Endo, M., and Koyama, T.: *J. Cryst. Growth* **32**, 335 (1976); Endo, M.: *Chemtech* **8**, 568 (1988)
8. Hirsch, A.: *The Chemistry of the Fullerenes*, Thieme, Stuttgart (1994)
9. Tutt, L.W., and Kost, A.: *Nature* **356**, 225 (1992)
10. Xie, Q., Perez-Cordero, E., and Echegoyen, L.: *J. Am. Chem. Soc.* **114**, 3978 (1992)
11. Guldi, D.M., and Martin, N. (Eds): *Fullerenes: From Synthesis to Optoelectronic Applications*, Kluwer Academic Publisher, Dordrecht (2002)
12. Thess, A., Lee, R., Nikolaev, P., Dai, H., Petit, P., Robert, J., Xu, C., Lee, Y.H., Kim, S.G., Rinzler, A.G., Colbert, D.T., Scuseria, G.E., Tomanek, D., Fischer, J.E., and Smalley, R.E.: *Science* **273**, 483 (1996)
13. Loiseau, A., Willaime, F., Demoncey, N., Hug, G., and Pascard, H.: *Phys. Rev. Lett.* **76**, 4737 (1996). For a review see Loiseau, A., et al.: *Carbon* **36**, 743 (1998)

14. Lee, R.S., Gavillet, J., Lamy, de la Chapelle, M., Cochon, J.-L., Pigache, D., Thibault, J., Willaime, F., and Loiseau, A.: *Phys. Rev. B Rapid Comm.* **64**, 121405 (2001)
15. Tenne, R., Margulis, L., Genut, M., and Hodes, G.: *Nature* **360**, 444 (1992)
16. Iijima, S., and Ichihashi, T.: *Nature* **363**, 603 (1993); Bethune, D.S., Kiang, C.H., de Vries, M.S., Gorman, G., Savoy, R., Vasquez, J., and Beyers, R.: *Nature* **363**, 605 (1993)
17. Journet, C., Maser, W.K., Bernier, P., Loiseau, A., Lamy de la Chapelle, M., Lefrant, S., Deniard, P., Lee, R., and Fisher, J.E.: *Nature* **388**, 756 (1997)
18. Maser, W.K., Munoz, E., Benito, A.M., Martinez, M.T., de la Fuente, G.F., Maniette, Y., Anglaret, E., and Sauvajol, J.-L.: *Chem. Phys. Lett.* **292**, 587 (1998)
19. Laplaze, D., Bernier, P., Maser, W.K., Flamant, G., Guillard, T., and Loiseau, A.: *Carbon* **36**, 685 (1998); Alvarez, L., Guillard, T., Sauvajol, J.-L., Flamant, G., and Laplaze, D.: *Appl. Phys. A* **70**, 169 (2000)
20. Farhat, S.: *J. Nanosci. and Nanotech.* (2003) in press
21. See, for example, Poretzky, A.A., Schittenhelm, H., Fan, X., Lance, M.J., Al-lard, L.F. Jr, and Geohagan, D.B.: *Phys. Rev. B* **65**, 245425 (2002); Scott, C.D., Arepalli, S., Nikolaev, P., and Smalley, R.E.: *Appl. Phys. A* **72**, 573 (2001)
22. Dorval, N., Foutel-Richard, A., Cau, M., Loiseau, A., Attal-Trétout, B., Cochon, J.L., Pigache, D., Bouchardy, P., Krüger, V., Geigle, K.P.: *J. Nanosci. and Nanotech.* (2003) in press
23. Cheng, H.M., Su Li, F., Pan, G., Pan, H.Y., He, L.L., Sun, X., and Dresselhaus, M.S.: *Appl. Phys. Lett.* **72**, 3282 (1998)
24. Colomer, J.-F., Bister, G., Willems, I., Konya, Z., Fonseca, A., Van Tendeloo, G., Nagy, J.B.: *Chem. Commun.* **14**, 1343 (1999); Colomer, J.F., Stephan, C., Lefrant, S., Van Tendeloo, G., Willems, I., Konya, Z., Fonseca, A., Laurent, C., and Nagy, J.B.: *Chem. Phys. Lett.* **317**, 83 (2000)
25. Dai, H., Rinzler, A.G., Nikolaev, P., Thess, A., Colbert, D.T., and Smalley, R.E.: *Chem. Phys. Lett.* **260**, 471 (1996)
26. Ren, Z.F., Huang, Z.P., Xu, J.W., Wang, J.H., Bush, P., Siegal, M.P., and Provencio, P.N.: *Science* **282**, 1105 (1998)
27. Franklin, N., and Dai, H.: *Adv. Mater.* **12**, 890 (2000)
28. Charlier, J.C., De Vita, A., Blase, X., and Car, R.: *Science* **275**, 646 (1997)
29. For a review, see Loiseau, A., Gavillet, J., Ducastelle, F., Thibault, J., Stéphan, O., and Bernier, P.: *C.R. Physics* (2003) in press
30. Tibbetts, G.G.: *J. Cryst. Growth* **66**, 632 (1984)
31. Saito, Y.: *Carbon* **33**, 979 (1995)
32. Gavillet, J., Loiseau, A., Journet, C., Willaime, F., Ducastelle, F., and Charlier, J.C.: *Phys. Rev. Lett.* **87**, 275504 (2001)
33. Amelinckx, S., Lucas, A.A., and Lambin, Ph.: *Rep. Prog. Phys.* **62**, 1471 (1999)
34. Kociak, M., Hirahara, K., Suenaga, K., and Iijima, S.: *Eur. Phys. J. B* **32**, 457-469 (2003)
35. Henrard, L., Loiseau, A., Journet, C., and Bernier, P.: *Eur. Phys. J. B* **13**, 661 (2000); Colomer, J.-F., Henrard, L., Lambin, Ph., and Van Tendeloo, G.: *Phys. Rev. B* **64**, 125425 (2001); *Eur. Phys. J. B* **27**, 111 (2002)
36. Wildöer, J.W.G., et al.: *Nature* **391**, 59 (1998)
37. Website of C. Dekker at Delft University (Netherlands)
38. Yakobson, B.I., and Smalley, R.E.: *La Recherche (Paris)* **307**, 50 (1998)

39. Smith, B.W., Monthieux, M., and Luzzi, D.E., *Nature* **396**, 323 (1998)
40. Guerret-Piécourt, C., Le Bouar, Y., Loiseau, A., and Pascard, H.: *Nature* **372**, 761 (1994)
41. Petit, P., and Loiseau, A. (Eds): *Nanotubes: Science and Applications*, C.R. Physics (Elsevier) (2003)
42. Derycke, V., Martel, R., Appenzeller, J., and Avouris, P.: *Nano Letters* **1** (9), 453 (2001)
43. Marty, L., Bouchiat, V., Naud, C., Chaumont, M., Fournier, T., and Bonnot, A.M.: *Nano Letters* **3** (8), 1115 (2003)
44. Heer, W.A.D., Chatelain, A., Ugarte, D.: *Science* **270**, 1179–1180 (1995)
45. Gomer, R.: *Field Emission and Field Ionization*, Harvard University Press, Cambridge, MA (1961), p. 155
46. Kasumov, A.Y., Deblock, R., Kociak, M., Reulet, B., Bouchiat, H., Khodos, I.I., Gorbatov, Y.B., Volkov, V.T., Journet, C., and Burghard, M.: *Science* **284** (5419), 1508–1511 (1999)
47. Choi, K.H., Bourgoin, J.P., Auvray, S., Esteve, D., Duesberg, G.S., Roth, S., and Burghard, M.: *Surf. Sci.* **462**, 195 (2000); Valentin, E., Auvray, S., Goethals, J., Lewenstein, J., Capes, L., Filoramo, A., Ribayrol, A., Tsui, R., Bourgoin, J.P., and Patillon, J.N.: *Microelectronic Engineering* **61–62**, 491 (2002)
48. Odom, T.W., Huang, J.-L., Kim, P., and Lieber, C.M.: *J. Phys. Chem. B* **104** (13), 2794 (2000)

## General References

49. Kittel, C.: *Introduction to Solid State Physics*, 7th edn. Wiley, New York (1996)
50. Dresselhaus, M.S., Dresselhaus, G., and Eklund, P.C.: *Science of Fullerenes and Carbon Nanotubes*, Academic Press, San Diego (1996)
51. Bernier, P., and Lefrant, S.: *Le carbone dans tous ses états*, Gordon & Breach Science (1997)
52. Saito, R., Dresselhaus, G., and Dresselhaus, M.S.: *Physical Properties of Carbon Nanotubes*, Imperial College Press (World Scientific, Singapore) (1998)
53. Dresselhaus, M.S., Dresselhaus, G., and Avouris, Ph.: *Carbon Nanotubes: Synthesis, Structure, Properties and Applications*, Topics in Applied Physics, Vol. 80, Springer, Berlin (2001)
54. Loiseau, A., Launois, P., Petit, P., Roche, S., Salvétat, J.P. (Eds): *Understanding Carbon Nanotubes: From Science to Applications*, Lecture Notes in Physics, Springer, Berlin (to be published in 2004)
55. Harris, P.J.F.: *Carbon Nanotubes and Related Structures. New materials for the Twenty-First Century*, University of Reading, UK (2002)
56. Tomanek, D., and Enbody, R.J. (Eds): *Science and Applications of Nanotubes (Fundamental Materials Research)*, International Conference (1999 East Lansing, Mich.) Nanotube '99 (Plenum Pub Corp) (2000)

## Nanowires

J.-C. Labrune and F. Palmino

A nanowire is generally defined as an object with a one-dimensional aspect in which the ratio of the length to the width is greater than 10 and the width does not exceed a few tens of nanometers. Today, this definition has been extended to atomic and molecular wires, which have proved to exhibit very interesting physical properties without necessarily having the geometrical characteristics defined earlier.

There are many potential applications for nanowires. However, nanowires can be divided into two categories:

- nanowires with direct applications determined by their physical properties,
- nanowires used as basic building blocks in more complex devices.

The direct applications mainly concern information storage, electronics and optonics. Examples are magnetic nanowires, light-emitting nanowires, or nanowires behaving as diodes. As elementary building blocks, nanowires are mainly used as electrical contacts, or as integral parts of components when they have semiconducting properties.

But whatever the application, nanowires are not yet an industrial product. Although there exist many ways of making them, the available methods do not yet combine mass production with very small dimensions. The subject of the day is not yet nanotechnology, but rather nanoscience. The observed physical phenomena, e.g., electrical conduction, heat conduction, mechanical properties, optical properties, magnetic properties, etc., have not all been explained or perfectly understood. But the stakes are high and it is essential to develop a complete mastery of nano-objects in general and nanowires in particular.

The aim in this chapter is to present the various techniques for making nanowires and to discuss transport phenomena within these quantum objects, at least in the case of electrical conduction.

## 9.1 Fabrication of Nanowires

There are two distinct approaches to the fabrication of nanostructures: the top-down approach and the bottom-up approach, one arising in the world of microelectronics and the other in the world of nanophysics. Although the overall aim may be the same, i.e., to produce nanostructures in the broad sense of the term, the way of going about it is totally different in the two cases. Huge resources are invested in developing techniques that can combine mass production and extremely high levels of resolution in the fabrication process, in order to achieve what is known as nanotechnology. Unfortunately, at the present time, no technique can really achieve this, and some prefer to stay with the term nanoscience.

In the top-down approach, one attempts to reduce the size of a complex object to the point where this scale reduction begins to alter the very principles it is based upon. The idea is seductive at first glance, but this method encounters major physical and technological difficulties when one attempts to go down to length scales of a few tens of nanometers using conventional lithographic methods. When these targets are reached, using emerging lithographic techniques, one has to face the problem of speed: the slow production rate is quite incompatible with the requirements of mass production.

The bottom-up approach is radically different, since it involves using atomic scale and nanoscale physicochemical phenomena to fabricate simple nanostructures in a spontaneous manner and in large quantities. The resources required in this case are considerably reduced since growth and assembly can be controlled in a single step, and in a natural and self-regulating manner. This control over crystal growth can be used to fabricate identical objects with the same properties, and at an incomparably lower cost. The disadvantage of this approach is that transistors, memory cells, and other components do not a priori form in a spontaneous way. The bottom-up approach thus requires the invention and study of new components compatible with this means of fabrication. This is why many specialists say that, if there is to be a revolution one day, it will necessarily be here.

Whatever approach is adopted, there are many methods for elaborating nanostructures and nanowires in particular. The choice of fabrication technique depends on several parameters, depending not only on the nano-object as such, but also on its function. Indeed, it is not enough to specify parameters like size, shape and character which define the nano-object. Other parameters are also essential when choosing the technique:

- the quantity of nano-objects to be fabricated,
- the spacing between the objects,
- their orientation in (2D or 3D) space,
- the nature of the substrate on which they are to be used,
- the manipulations required after fabrication.

Apart from lithographic techniques used in the microelectronics industry which we shall not be concerned with here, the work currently being carried out in the field of nanowire fabrication can be classified as follows:

1. Using a top-down approach:
  - a) conventional high resolution lithographic techniques (electron beam, extreme UV, and X ray),
  - b) alternative lithographic techniques:
    - nanoimprinting,
    - nanomoulding,
    - lithography by near-field microscopy,
2. Using a bottom-up approach:
  - a) self-assembly techniques,
  - b) VLS synthesis,
  - c) use of porous matrices.

## 9.2 The Top-Down Approach

Two serious problems face the top-down approach. These are raised by the limits of photolithographic techniques, which are the only ones able to mass produce submicron components. These processes provide an extremely effective way of fabricating structures with dimensions around a hundred nanometers, but they are faced with almost insoluble problems when one attempts to increase their resolution beyond this point. UV and X-ray lithography are currently being investigated with the aim of going well below the critical 100-nm threshold that marks the gateway to nanotechnology. However, the prohibitive costs involved leave little hope for rapid application of these new forms of technology. As far as electron beam lithography is concerned, it is at present the only way of etching nanostructures on the scale of a few nanometers. Unfortunately this technique is still incompatible with mass production. It should be noted that this problem, the confrontation between nanometric resolution and mass production, recurs throughout the field of nanotechnology. This is why emerging forms of lithography known as soft lithography have appeared on the scene since the 1990s.

### 9.2.1 Soft Lithography

The basic idea is to make low cost replicas of nanostructures that have been fabricated using complex and costly techniques, such as electron beam lithography, for example. To this end, one deposits substances with extremely faithful replication properties, which harden and copy the patterns in reverse relief. The best known are polydimethylsiloxane or PDMS and to a lesser extent polymethyl methacrylate or PMMA (thermoplastic). This produces

a stamp that can be used several times over and hence reproduce nanostructures on a large scale and at low cost [1, 2]. It is worth noting that the nanostructures produced as yet have not been able to achieve the complexity of structures obtained using standard lithographic techniques. The two main techniques available today which use this type of stamp are nanoimprinting and nanomoulding.

### **Nanoimprinting**

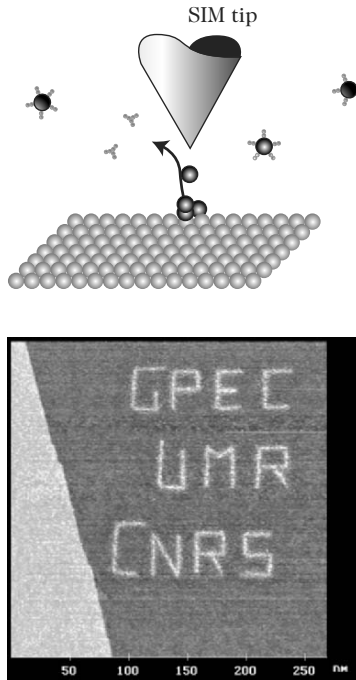
Nanoimprinting really does use the replica as a stamp. PDMS is an elastomer which combines several interesting properties: it has a suitable Young's modulus for moulding, it is non-toxic, it is available for industrial use, it is optically transparent around 300 nm, it is chemically inert, and its interface free energy can be considerably modified. PDMS is extremely hydrophobic, but it can be rendered hydrophilic by plasma treatment. The possibility of considerably modifying its surface free energy means that it can be coated with a monolayer of molecules which can then, under certain conditions, be retransferred to a surface. The best known example is the transfer of a monolayer of thiols onto a gold surface, although deposits onto Ag or Cu are also possible [3–5]. To do this, the PDMS patterns are inked up using a solution containing thiols. When the stamp is pressed against the gold surface, there is a strong reaction between the thiols and the surface, whereupon the thiols detach themselves from the stamp to form a self-assembled layer which constitutes a faithful copy of the stamp. This kind of technique can produce patterns with an accuracy down to a few tens of nanometers.

### **Nanomoulding**

The two main variants of nanomoulding are nanoembossing and capillary moulding [6–11]. In the first case, the idea is to impress the stamp upon a liquid polymer, or a polymer that has been softened by heating. When it cools, the polymer solidifies and adopts the shape of the mould. In the second case, the liquid polymer is deposited at the entrance to cavities that have been left open between the 'stamp' and the plane surface on which it has been deposited. The liquid polymer then flows by the capillary effect between the surface and the 'stamp' to reproduce the required patterns when it solidifies. Patterns smaller than 30 nm wide have already been obtained by this method. Once the pattern has been transferred to the surface, metallic nanowires of very small dimensions can be created by traditional techniques for making openings, metallic deposit and lift-off.

#### **9.2.2 Near-Field Lithography**

Near-field lithography uses the tip-sample interaction of a near-field microscope to etch, deposit, move atoms, control chemical reactions, and so on.



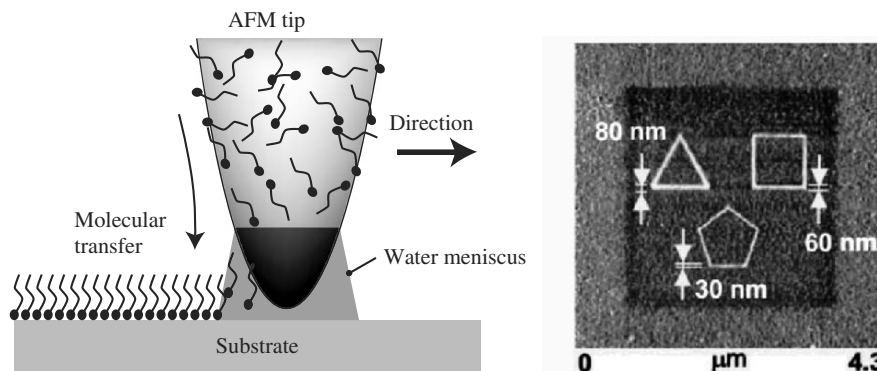
**Fig. 9.1.** *Upper:* Diagrammatic representation of the decomposition of a gas under the tip and adsorption of the element on the surface. The decomposition voltage can be pulsed or continuous. *Lower:* STM image obtained by decomposition of germane on Si(111), forming nanoletters in germanium. Courtesy of F. Thibaudau (unpublished)

It can thus modify surfaces on the atomic scale. This use of near-field microscopy provides the best resolutions available today in the field of nanowire fabrication. However, it is also very difficult to implement and cannot be used for large scale production. The most developed techniques in this field are described here.

### STM-Assisted CVD

This is the method used to fabricate silicon or metal nanowires for nanocomponents [20–22]. It involves using the energy of electrons moving between the tip and the sample during an STM experiment in ultrahigh vacuum to decompose a precursor gas, e.g., Ni(Co), Fe(CO)<sub>5</sub>, SiH<sub>2</sub>Cl<sub>2</sub>, AuCl(PF<sub>3</sub>), and adsorb one of the constitutive elements of this gas onto the surface. The tunnel current is one-way and decomposition can be very accurately localised under the tip. The tip is then moved along preprogrammed directions to lay down structures a few nanometers across by nanolithography (see Fig. 9.1). Apart from its resolution, the advantage with this technique is that one can use different gas compositions in order to choose (at least to some extent) the composition of the nanowire.





**Fig. 9.2.** *Left:* Transferring thiols onto a surface. *Right:* AFM image of thiol nanowires. Taken from [13]

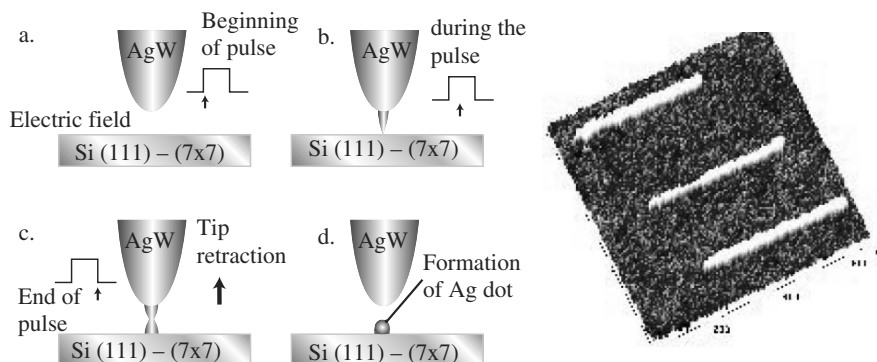
### Chemical Reactions Induced by the Tip

Another interesting illustration of near-field microscopy is provided by the work of Legrand et al. [23] at IEMN (Lille, France), who fabricated silicon nanowires using a biased AFM tip. A potential is applied between the tip and sample, causing local oxidation of a hydrogenated silicon surface under the tip. A wet etch by KOH then destroys the unoxidised regions, creating the silicon nanowires.

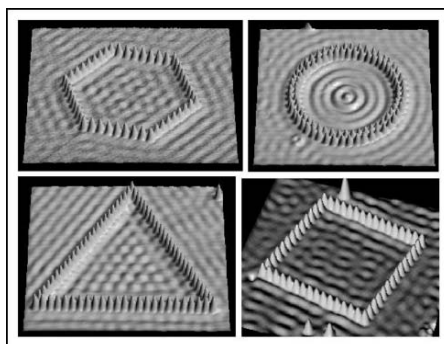
### Matter Transfer Using the Tip

The idea here is to transfer the atoms or molecules coating the tip of a near-field microscope onto the surface. The control one has over this transfer of matter depends on the tip-sample interaction and on the type of matter to be moved. Today, knowhow acquired in this field is sufficient to carry out this kind of operation with great precision, using not only the STM technique, but also the AFM technique. The best known example is certainly the thiol film adsorbed onto an AFM tip and transferred to a gold surface (see Fig. 9.2) [12–14].

The transfer is achieved via a nanodroplet of water which plays the role of a bridge between the tip and the sample surface. The driving forces come from surface tension generated by the droplet. The thiols, insoluble in water, fix themselves very locally on the surface in a self-organised layer. There are other examples which use an STM tip for local transfer of silver, gold, platinum, etc. (see Fig. 9.3) [15–19]. By applying a voltage pulse between the tip and sample, the electric field increases significantly and this leads to diffusion of silver atoms from the tip to the surface. There again, a very high level of accuracy is achieved (about 20 nm), but the quantities produced are well below the mark.



**Fig. 9.3.** *Left:* The various stages in a transfer of matter between the tip and sample. *Right:* STM image of platinum nanowires made by field emission STM. Courtesy of Houel et al. [15]



**Fig. 9.4.** Geometric figures obtained using iron atoms displaced by an STM tip on a copper surface at low temperature [26]

### Manipulation of Atoms and Molecules

The fabrication of atomic scale 1D structures is now well established [24, 25]. Made by lining up atoms along a crystal axis, these structures constitute the smallest ‘nanowires’ ever fabricated. Indeed, one may ask whether the appellation is still appropriate. Figure 9.4 shows an atom-by-atom organisation of iron atoms on a copper surface at very low temperature.

On this length scale, stability (or lifetime) problems arise for the nanowire, given the very small number of atoms involved and the very low temperatures required for such structures. The techniques used to make them are also extremely complex and costly. The atoms are displaced and positioned on the surface by means of the attractive and repulsive forces between the tip and sample. By adjusting the current/voltage parameters, the tip can be used to push, drag, catch or release atoms adsorbed on the surface. The result is a kind of atomic scale construction game which can be implemented on either

semiconducting or metallic surfaces. Here again, surface effects dominate the position and stability of the atoms. Hence, an atom can only remain on a given site if it corresponds to a state of lowest energy. Otherwise, if the temperature is high enough, it will wander from site to site in search of this energy minimum, thereby destroying the nanowire. The most favourable conditions for stabilising the system involve working at very low temperatures. The first experiments were carried out by Eigler et al. [26], who displaced Xe atoms on Ni(100) at 4 K. The technical aspects and fabrication time are well removed from the requirements of industry.

## 9.3 The Bottom-Up Approach

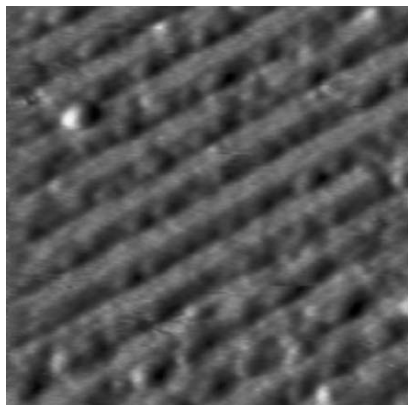
### 9.3.1 Self-Assembly on a Surface

Research into the possibility of building nanowires by self-assembly began only recently. It was triggered by the discovery and development of near-field microscopy techniques (STM and AFM) in the 1990s. Using these two new instruments, considerable progress has been made in understanding surfaces and interfaces, to the point where extremely selective and tightly controlled growth becomes feasible. These experiments are usually carried out in ultra-high vacuum conditions, even if the electrochemical channel can also lead to very good results. Self-assembly (also known as self-organisation) is a very interesting phenomenon indeed for those who seek to create on a surface a large number of perfectly ordered objects, with simple shape and the same size. The basic idea is to use a surface which exhibits a selective and strong adsorption in a highly localised and periodic manner. Sites of preferential adsorption serve as anchoring points for growing nanostructures. Periodicity can be obtained in different ways, over a region from 1 to 100 nm across. At the present time, three self-assembly techniques have been devised for nanowire fabrication:

- use of periodic surface reconstructions of 1D type,
- use of a stress field induced by lattice mismatch during epitaxial growth,
- use of step edges on a vicinal surface.

#### Prepatterned Surfaces

The use of surfaces with a crystal reconstruction of 1D type (reconstructed surfaces) is one of the most delicate techniques to implement. One of the reasons for this is that it is extremely difficult to predict the crystal organisation of a surface because only the structures with the lowest energy can occur. It is generally accepted today that what determines the most stable structure results from the competition between minimisation of the electron binding energy and minimisation of the elastic stress energy. The solution to



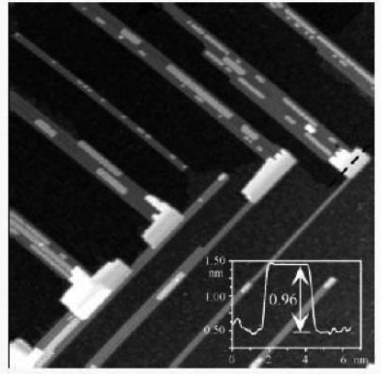
**Fig. 9.5.** STM image ( $7\text{ nm} \times 7\text{ nm}$ ) of lead nanowires self-organised on an SmSi(111) interface. Courtesy of Palmino et al. (unpublished)

the problem is clearly not trivial, and all the more so in that any further adsorption for the purpose of creating the nanowire will induce a change in its energy and possibly a change in its crystal structure, which may destroy its 1D character. The problem here is thus twofold: to create 1D surface reconstructions [27–30] and to maintain this one-dimensionality [31, 32] during adsorption and growth. One can then speak of prepatterned surfaces where deposited atoms must replicate the pattern on the surface. Given these constraints, very few systems exhibit this functionality (see Fig. 9.5).

### Vicinal Surfaces

As we have seen, it is not easy to obtain 1D prepatterned surfaces. One way around this problem is to use natural defects which are in principle easier to obtain. The best known is the step edge. It has the advantage of being highly reactive and one-dimensional, and extending over large distances. Furthermore, it occurs on every type of surface. The number and distribution of steps on the sample surface are mainly determined by the crystal orientation, the heat treatment undergone by the crystal during preparation, and the interactions between steps (lamellar materials such as graphite are not included in this category). As the first two parameters are generally well controlled, the distance between steps can be relatively easily adjusted.

There are many examples of the use of step edges to create nanowires. We may cite the decoration of step edges by electrodeposition of  $\text{MoO}_2$  on HOPG (highly oriented pyrolytic graphite) to obtain molybdenum nanowires [33]. Other studies have produced nanowires by selective electrodeposition of metals on vicinal H-Si(111) surfaces. In ultrahigh vacuum, we may cite the fabrication of Ag and Au wires on vicinal surfaces of platinum and silicon [34–37]. The great interest of silicon surfaces is the excellent level of control of the surface structure on large scales and on the atomic scale.



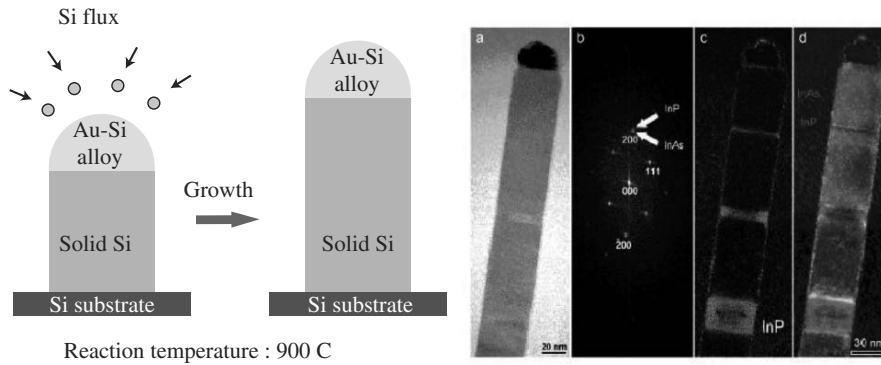
**Fig. 9.6.** STM image of dysprosium silicide nanowires on Si(100) induced by parameter mismatch. From J. Nogami

### Parameter Mismatch

The lattice parameter mismatch between two solid phases is often used during crystal growth to generate nanostructures by the effects of elastic stress relaxation during Stranski–Krastanov-type growth. When this parameter mismatch is greater than 2%, the stress energy of the deposited film is released by island formation. The shape of the islands formed in this way depends on the crystal symmetry of the substrate, but also on the parameter mismatch occurring in certain directions. For example, in the case of Ge/Si(001) and InAs/Ga(001) interfaces, 2D or 3D islands appear on the surface (not always uniformly as far as shape is concerned). In order to make nanowires, the epitaxial layer must exhibit a large parameter mismatch in one direction and a low one in a perpendicular direction. Among the few systems which do exhibit such characteristics, we may cite certain silicon/rare earth interfaces such as Ho/Si(100), Gd/Si(1000) or Dy/Si(100) (see Fig. 9.6) [38, 39].

#### 9.3.2 VLS Synthesis

The VLS (vapour–liquid–solid) technique consists in growing nanowires from a molten droplet [40, 41]. This droplet is created in a high temperature reactor (around 900°C) and fed either by a vapour phase produced from a target bombarded by a laser, or by CVD, or both at once. The method is particularly well suited to the growth of semiconductors and metal alloys for which the phase diagrams can be perfectly controlled. Nanowires of diameter 10 nm and length several microns can be fabricated in this way. With this technique, it is also possible to vary the composition of the interior of the nanowire by controlling the composition of the vapour phase arriving on the droplet. It is then possible to make doped or undoped semiconducting nanowires (Si, GaAs, InP, GaN, etc.) or multilayer metallic nanowires with magnetic properties, and even to control to a certain extent the composition of the outer part of the wire. The growth of the wire is directly linked to the Gibbs–Thomson effect



**Fig. 9.7.** VLS fabrication method. Image of a multilayer InAs/InP nanowire with diameter 30 nm. Courtesy of Karaguchi et al. [41]

which shows that the smaller the diameter, the slower the growth will be (see Fig. 9.7).

### 9.3.3 Use of Porous Matrices

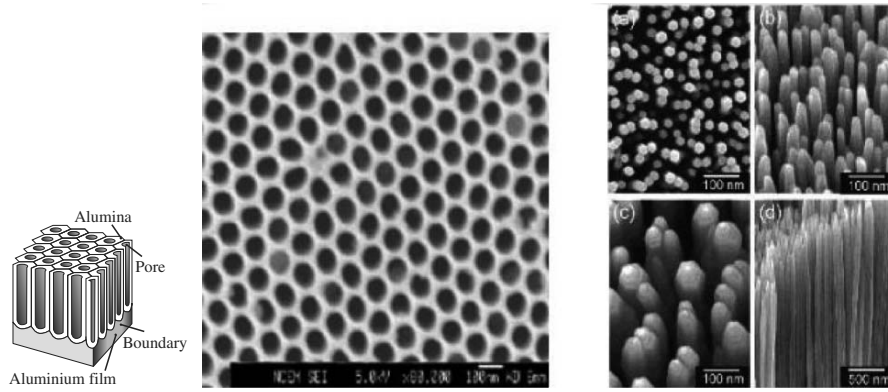
In this method, a porous lattice or matrix, known as the template is filled with matter. In general, this 3D aluminium oxide matrix, fabricated on a silicon substrate, is composed of perfectly calibrated vertical nanopores, which may be as much as a few tens of microns long. Due to the geometry of the system, a high concentration of nanowires is obtained with orientation perpendicular to the surface at the end of the fabrication process, i.e., after the matrix has been dissolved. This feature of the orientation contrasts with the results of self-assembly techniques. The pores can be filled in different ways: electrochemical, high-pressure injection, or evaporation. However, the electrochemical approach seems to be the best suited, because particularly long, continuous nanowires can be obtained with either metallic or semiconducting materials (see Fig. 9.8) [42–44].

## 9.4 Electrical Conduction in Nanowires

The electrical conductance is a fundamental quantity encountered in every study of electrical transport phenomena in macroscopic systems, mesoscopic systems, and nanowires.

A macroscopic conductor with cross-sectional area  $S$  and length  $L$ , subjected to a potential difference  $V$ , carries a current  $I$ . Ohm's law says that  $I = G/V$  and the conductance  $G$  is given by  $\sigma S/L$ , where  $\sigma$  is the conductivity of the conducting material.

On the mesoscopic scale, this law no longer applies. A quantum description is required. The system is assumed to be coherent and its electrons are



**Fig. 9.8.** *Left and center:* Diagram and MEB image of an  $\text{Al}_2\text{O}_3$  porous matrix. Courtesy of Sands et al. [43]. *Right:* ZnO nanowires. Courtesy of Park et al. [44]

described by wave functions with well defined phase. On this scale, there are different electron transport regimes. One must compare the geometrical dimensions of the conductor with the mean free path  $L_m$  and the phase coherence length  $L_\varphi$  of the electrons. The latter is defined as the average distance separating two inelastic collisions and during which the electron retains its phase memory. For example, for gold at 1 K,  $L_\varphi$  is close to one micron. One thus speaks of the ballistic regime when  $L_m > L$  and the diffusive regime when  $L_m \ll L$ . Any mesoscopic system will have  $L_\varphi < L$  and its Fermi wavelength  $\lambda_F$  will be very small compared with its geometrical characteristics.

#### 9.4.1 Electrical Contacts

In order to carry out electrical measurements on a mesoscopic sample or a nanowire, connections have to be made. In the case of a nanowire, the contacts impose specific constraints. Indeed, there are several orders of magnitude between the size of the wire and the size of the link with the measuring apparatus. Moreover, techniques for fabricating contacts and nanowires are generally very different.

At the end of the 1980s, the first conductance measurements were made on atomic scale contacts using near-field microscopy (STM) [45]. The quantisation of the conductance  $G$  was observed in semiconducting heterojunctions [46,47], and confirmed a little later in experiments using break junctions [48]. These experimental results heralded a great deal of theoretical activity with regard to such a quantisation and concerted attempts at numerical simulation. For this purpose, one requires theories and models for electrical transport in mesoscopic conductors [49, 50].

### Basic Model

The basic model for theoretical study in this context considers the mesoscopic sample joined to two reservoirs by two conducting connections, assumed perfectly conducting, through which electrons can be injected into the sample. The reservoirs are at temperatures  $T_1$  and  $T_2$  and have chemical potentials  $\mu_1$  and  $\mu_2$ , whose difference is in our case proportional to the potential difference set up between the two reservoirs, which play the role of electrodes. It is assumed that all inelastic scattering is limited to the electrons in the reservoirs and that phase coherence is maintained within the sample. Along the perfect conductors, plane waves are associated with electrons propagating in the longitudinal direction. Due to confinement, the transverse moment is quantised. Let  $N_1$  and  $N_2$  denote the numbers of transverse electronic modes of electrons in the two connections, which behave as waveguides. The energy distribution of these electrons depends on  $\mu_1$  and  $\mu_2$  and obeys the Fermi distribution  $f(\varepsilon)$ . The outgoing electrons are transmitted to the reservoirs with probability equal to unity.

### Single-Mode Sample

In the simple case of a perfect ballistic sample, with only one occupied mode, there is a current due to the population difference of the mode propagating to the left and to the right. Taking spin degeneracy into account, the current is given by

$$I = \frac{2e}{h} \int [f_1(\varepsilon) - f_2(\varepsilon)] d\varepsilon . \quad (9.1)$$

At 0 K, the Fermi distributions  $f_1(\varepsilon)$  and  $f_2(\varepsilon)$  reach their asymptotic values (1 then 0) on either side of the energies  $\varepsilon_1 = \varepsilon_F + eV/2$  and  $\varepsilon_2 = \varepsilon_F - eV/2$ , respectively. Finally, we obtain the conductance  $G = I/V$  which has the universal value

$$G = G_0 = 2e^2/h = 0.77 \mu\text{S} , \quad (9.2)$$

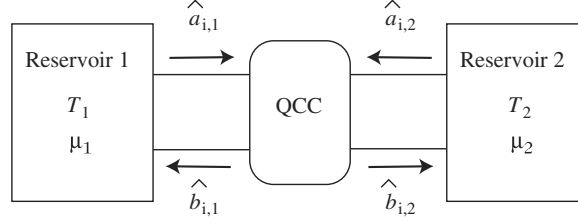
where  $S$  is the siemens, SI unit of electrical conductance.

### Landauer Theory and Linear Response

The Landauer theory [51–53], further developed by Büttiker [54, 55], relates the conductance to the transmission and reflection probabilities of the electron waves arriving at the quantum coherent conducting sample. The theory is based on the scattering matrix

$$\hat{S} = \begin{pmatrix} \hat{s}_{11} & \hat{s}_{12} \\ \hat{s}_{21} & \hat{s}_{22} \end{pmatrix} = \begin{pmatrix} \hat{r} & \hat{t}' \\ \hat{t} & \hat{r}' \end{pmatrix} . \quad (9.3)$$





**Fig. 9.9.** A quantum coherent conductor (QCC) is joined to two reservoirs (electrodes), with chemical potentials  $\mu_1$  and  $\mu_2$  and temperatures  $T_1$  and  $T_2$ , by two connections assumed to be perfectly conducting. The operators corresponding to the  $i$ th propagation mode can be decomposed into eight creation and annihilation operators, incoming and outgoing with respect to the conductor

Each entry  $\hat{s}_{\alpha\beta}$  is a matrix whose components  $(\hat{s}_{\alpha\beta})_{ij}$  express the transmission ( $\alpha \neq \beta$ ) or reflection ( $\alpha = \beta$ ) of the waves. The theory also introduces operators  $\hat{a}$  associated with electrons coming into the sample and operators  $\hat{b}$  for those leaving the sample (see Fig. 9.9).

Different operators correspond to creation and annihilation of an electron in the  $i$ th mode coming from reservoir  $\alpha$  ( $\alpha = 1$  or  $2$ ) and entering the sample, denoted by  $\hat{a}_{\alpha,i}^\dagger$  and  $\hat{a}_{\alpha,i}$ , respectively. For an electron leaving the sample and returning to the reservoir  $\alpha$ , these operators are denoted by  $\hat{b}_{\alpha,i}^\dagger$  and  $\hat{b}_{\alpha,i}$ . The scattering matrix relates these operators by

$$b_{i,\alpha} = \sum_{j\beta} (\hat{s}_{\alpha\beta})_{ij} a_{j,\beta}.$$

We may also write

$$\langle \hat{a}_{i,\alpha}^\dagger(\varepsilon) a_{j,\beta}(\varepsilon) \rangle = \delta_{ij} \delta_{\alpha\beta} f_\alpha(\varepsilon),$$

where  $f_\alpha(\varepsilon)$  is the Fermi function of the reservoir joined to the connection  $\alpha$ .

Assuming an ideal coupling between reservoirs and connections, the current of the  $i$ th mode in the connection  $\alpha$  can be written

$$I_{i,\alpha} = \frac{2e}{h} \int_{-\infty}^{+\infty} \left[ \langle \hat{a}_{i,\alpha}^\dagger(\varepsilon) a_{i,\alpha}(\varepsilon) \rangle - \langle \hat{b}_{i,\alpha}^\dagger(\varepsilon) b_{i,\alpha}(\varepsilon) \rangle \right] d\varepsilon. \quad (9.4)$$

Summing over all modes  $i$ , we obtain the total current  $I_1$  in the left-hand connection:

$$I_1 = \frac{2e}{h} \int_{-\infty}^{+\infty} \left[ (N_1 - R_{11}) f_1 - T_{12} f_2 \right] d\varepsilon. \quad (9.5)$$

The reflection and transmission coefficients,  $R_{11}$  and  $T_{12}$ , respectively, are given by the traces  $\text{Tr}(\hat{r}^\dagger \hat{r})$  and  $\text{Tr}(\hat{t}^\dagger \hat{t})$  of the elements of the matrices  $\hat{S}^\dagger$  and

$\hat{S}$ , respectively. Conservation of the total current requires the matrix  $\hat{S}$  to be unitary, i.e.,  $\hat{r}^\dagger \hat{r} + \hat{t}^\dagger \hat{t} = \hat{1}$  or  $R_{11} + T_{12} = 1$ . The current  $I_1$  then reduces to

$$I_1 = \frac{2e}{h} \int_{-\infty}^{+\infty} T_{12}(f_1 - f_2) d\varepsilon. \quad (9.6)$$

The linear response of the nanowire is defined in the vicinity of the equilibrium state  $I_1 = 0$  obtained when  $\mu_1 = \mu_2$ . For small variations in the applied voltage  $V$  the variation  $\delta I_1$  of the current is proportional to  $V = (\mu_1 - \mu_2)/e$  and hence,

$$G = \frac{\delta I_1}{(\mu_1 - \mu_2)/e} = \frac{2e}{h} \int_{-\infty}^{+\infty} T_{12} \left( -\frac{\partial f}{\partial \varepsilon} \right) d\varepsilon. \quad (9.7)$$

For  $T = 0$  K, we obtain

$$G = \frac{2e^2}{h} T_{12}. \quad (9.8)$$

As the trace of a Hermitian matrix,  $T_{12}$  has real eigenvalues  $\tau_i$  ( $i \in \{1, N_1\}$ ) with  $\tau_i \leq 1$  for all  $i$ . In the basis of eigenvectors (also called eigenchannels) of  $\hat{r}^\dagger \hat{r}$  and  $\hat{t}^\dagger \hat{t}$ , the multimode transport problem may be treated as a superposition of independent single-mode problems leading to a conductance expressed in the form of a sum:

$$G = \frac{2e^2}{h} \sum_i \tau_i = G_0 \sum_i \tau_i. \quad (9.9)$$

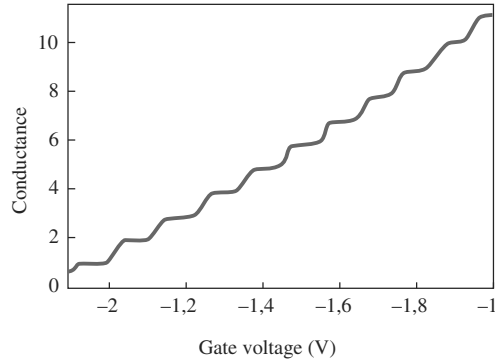
The modes to be taken into account, i.e., those with nonzero coefficient  $\tau_i$ , are determined by the sample itself, and not by the connections to the reservoirs. How many there are depends on the nature of the sample material. For most metals, it is limited to a number between 1 and 3.

### Experimental Corroboration

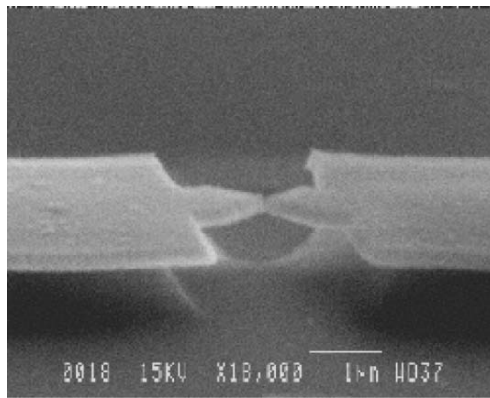
The Landauer linear response theory was confirmed in 1988 in an experiment using GaAs–AlGaAs heterojunctions [46] with a thin conducting film. The conductance was observed to vary by steps of height  $G_0$  when the dimensions of the narrow constriction formed at the level of the conducting film were varied by applying a variable negative bias at the gate. The experiment was carried out at a temperature very close to 0 K (see Fig. 9.10).

At the beginning of the 1990s much experimental and theoretical work was carried out on atomic scale metallic contacts. Two techniques made these studies possible: break junctions [48] and near-field electron microscopy [45].

The break junction approach illustrated in Fig. 9.11 consists in stretching a wire (or a film deposited on a substrate) and breaking it by steadily twisting



**Fig. 9.10.** Conductance  $G$  as a function of the bias applied at the gate of a GaAs–AlGaAs junction at 0.6 K. Note that, after subtracting the resistance of the connecting wires, plateaus are located at multiples of  $G_0$ . Taken from van Wees et al. [46]



**Fig. 9.11.** Break junction made by lithography, constituted by a gold film of thickness 20 nm in contact with an aluminium film of thickness 400 nm. Taken from Scheer et al. [56]

it. The contact is then reestablished by very carefully bringing together the two pieces of wire.

Something similar can be done with the tip of an STM when it is brought into contact with a surface, thereby moving from the tunneling regime to a metallic contact. This contact, which only involves a very small number of atoms, is reflected by a sudden change in the conductance. When the tip is withdrawn, one observes plateaus in  $G$ , located at multiples of the universal value  $G_0$ .

For many monatomic metals (Au, Ag, Cu, Li, Na, K, etc.), a last plateau is indeed observed at the value  $G_0$ . However, for other elements, this is not the case. Various simulation techniques have been used, based upon simplifying

hypotheses concerning the forces acting on the electrons, the number of atoms included in the calculations, dynamical and energy considerations, and so on. We have seen that the transmission probabilities  $\tau_i$  for the channels  $i$  appear in the expression for the conductance. For a single-atom contact, the number of eigenchannels is fixed by the number of valence orbitals of the atoms. For example, for a metal of type *sd*, 5 or 6 channels can be partially opened ( $\tau_i < 1$ ). We may conclude that quantisation of the conductance in multiples of  $2e^2/h$  is related to the size of the contact and is only relevant for monovalent atoms.

### Other Formalisms

Some theoretical studies and numerical simulations use Green functions, a more powerful tool than the scattering matrix  $\hat{S}$ . Indeed, they yield the response  $A(x)$  at any point  $x$  inside or outside a system following an excitation  $A'(x')$  at some other point  $x'$ . These functions are useful for studying transport phenomena when electron–electron or electron–phonon interactions are taken into account.

In a crystal, the current density  $\mathbf{j}$  is related to the electric field vector  $\mathbf{E}$  by the local Ohm’s law:  $\mathbf{j} = \sigma \mathbf{E}$ , where  $\sigma$  is the conductivity which depends on the angular frequency  $\omega$  of the applied electric field. The Kubo formalism [57] can be used to express the conductivity for a mesoscopic conductor, relating it to the conductance using retarded and advanced Green functions  $G^{\text{r,a}}(\mathbf{r}, \mathbf{r}', E)$  whose main features are described in [49]. The transmission coefficient  $T_{12}$  in the Landauer relation (9.8) can then be identified with the quantity

$$T_{12} = \hbar^2 \sum_{ij} v_i v_j \left| G_{ij}^{\text{a}}(x, x', E) \right|^2, \quad (9.10)$$

where  $v_i$  is the speed in channel  $i$ , equal to  $\hbar k_i/m_e$ , and  $m_e$  is the mass of the electron. Such a relation can be exploited in numerical simulations using the tight-binding model [58].

### Numerical Simulations

Some simulations have been able to explain the mechanisms governing the formation, evolution, and conduction of atomic scale contacts. For certain materials, perturbations to the quantisation of the conductance have also been explained. Several mechanisms have been elucidated using molecular dynamics techniques, sometimes coupled with the tight-binding or free-electron models [58]. It has thus been possible to show the following:

- when a tip is brought towards a surface to establish a contact, the atoms situated at the interface will make an irreversible jump of several tenths of a nanometer in a time of about a picosecond,

- when the tip is retracted, there is a plastic deformation of the contact resulting in a non-adiabatic constriction involving various atomic rearrangements,
- this constriction, although totally crystalline, contains a number of defects (gaps, local disorder, surface roughness) which perturb the quantisation of the conductance,
- for contacts involving more than two atoms, the conductance plateaus are no longer situated at integer multiples of  $G_0$ .

Mechanisms for partial opening or total closure of conduction channels built into certain studies have been able to support the experimental observations. Recent *ab initio* methods [59, 60] include a detailed atomic-level description of the electrodes and the influence they have on the contact conductance. For example, they show that, in the case of aluminium, one must take into account the geometry of the whole system in order to interpret conductance measurements, something which has not been observed for the metal most widely used by research scientists, i.e., gold.

#### 9.4.2 Incoherent Transport

Despite its many successes, the diffusive approach to conduction is not beyond criticism. This theory is only valid if one neglects the inelastic scattering processes affecting the electrons in the mesoscopic wire or nanowire, which is effectively what happens when one assumes that propagation is coherent. It is thus only strictly valid when  $T = 0$  K, although it remains applicable at very low temperatures, and for electrons with energies close to the Fermi energy. Although the expression (9.6) giving the current  $I_1$  in the perfect connection is a priori independent of the applied bias, there is no reason for thinking that the transmission coefficients  $\tau_i$  will also be independent of this bias and the energy. A great deal of work has been carried out to determine the profile of the potential along the wire, using self-consistent theories and Green functions [61]. There are cases where transport is completely dominated by electrical effects leading for example to Coulomb blockade [62]. Other examples of incoherent transport are encountered in molecular nanowires.

#### 9.4.3 Atomic Chains and Molecules

Current methods of nanofabrication can be used to obtain truly 1D systems via self-assembly [63], near-field lithography [64], the break junction technique [65], or conjugated organic molecules [66].

Over the last five years, many theoretical studies have been brought to bear on electrical conduction in atomic chains and in individual molecules. The first used 1D models [67–70], or 2D models [71, 72], and sometimes 3D models [73]. The latter take into account the reservoirs to which the molecules are joined. Through these studies, certain points have been brought out, such

as the importance of the electronic and chemical interaction at the level of the contact with the reservoirs, the importance of the relative position of the Fermi level of the electrode and the energy levels at the end of the wire, and the importance of the gap between the levels of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). It is in the latter energy interval that transport can occur by the tunnel effect, for low biases and before Coulomb blockade.

The above theories assume elastic scattering of the electron in a rigid atomic lattice. However, in many situations, classical inelastic interactions of type electron–electron or electron–phonon cannot be neglected in such wires. These interactions can have important consequences such as the formation of a Luttinger liquid [74] in the case of electron–electron interactions [75] or a Peierls transition [76] in the case of electron–phonon interactions, in which an initially metallic system becomes semiconducting.

The latest research appeals to new ideas. One such theory suggests that an electron loses its identity when injected into such a system because, like all other such electrons, it must follow the same path in the wire or molecule. It divides into two quasi-particles: the spinon which carries the spin, without the charge, and the holon which carries the positive charge of a hole, without its spin [77]. Another theory [78] suggests that the electron–phonon interaction creates a virtual polaron which is a quasi-particle corresponding to a lattice distortion around the injected charge. At low temperatures and for molecular chains built with alternating single and double bonds (we speak of dimerisation), transport is governed by polaron propagation alone. For certain boundary conditions, the chain may contain a topological defect known as a soliton. At room temperature, transport is then a mixture of coherent processes (electron–phonon coupling) and incoherent processes (lattice fluctuations).

Other very recent work has concerned incoherent charge transport through molecular wires [79,80] or charge blocking states [81,82]. Electrical conduction measurements have been made on gold atomic wires of variable length, where only a few atoms are involved [83], obtained in an STM at low temperature, and on individual molecules [84]. In the first case, the STM is used to carry out conductance measurements on the atomic wire. A conductance of  $G_0$  is observed when the interelectrode potential difference  $V$  is zero, corresponding to a ballistic regime. It decreases as soon as the voltage reaches a threshold value  $V_t$  leading to emission of a phonon. The frequency of this phonon depends on the stress in the wire. The energy exchange between electrons and phonons is accompanied by heating of the atomic chain when phonons are emitted ( $V > V_t$ ) and cooling when phonons are absorbed ( $V < V_t$ ).

## 9.5 Conclusion

The high level of activity in the study of nanowires and nanotubes reflects the promise they hold for fields like optronics, electronics, and computing. This kind of research requires interdisciplinary skills borrowing from materials science, solid state physics, chemistry, optics, biology, and others. The many and varied fabrication processes available today can create insulating, metallic, and doped or undoped semiconducting nanowires. There are already a wide range of applications including luminescent diodes, logic inverters, field effect transistors, gas sensors, and detectors of biological compounds obtained by binding sensor molecules onto the surface of a nanowire [85]. Memory cells have also been built using nanowires assembled into a crisscross array [86]. At the present time, the main challenge lies in fitting together large numbers of these wires, and in improving reliability.

From the theoretical standpoint, a lot of work is in progress to understand the different phenomena encountered experimentally. Although the study of electrical conduction discussed in this chapter is of prime importance, other investigations are equally relevant, e.g., those concerning nanomagnetism, spin electronics, thermal transport, superconductivity, and so on. The interested reader is encouraged to read other chapters in the present book and the review [58].

## References

1. Pease, R.F.W.: *J. Vac. Sci. Technol. B* **10**, 278–285 (1992)
2. Zhao, X.-M., Xia, Y., and Whitesides, G.M.: *J. Mater. Chem.* **7**, 1069–1074 (1997)
3. Kumar, A., Biebuyck, H., and Whitesides, G.M.: *Langmuir* **10**, 1498–1511 (1994)
4. Xia, Y., Kim, E., and Whitesides, G.M.: *J. Electrochem. Soc.* **143**, 1070–1079 (1996)
5. Moffat, T.P., and Yang, H.: *J. Electrochem. Soc.* **142**, L220–222 (1995)
6. Xia, Y., and Whitesides, G.M.: *Soft Lithography*, *Angew. Chem. Int. Ed.* **37**, 550–575 (1998)
7. Xia, Y., McClelland, J.J., Gupta, R., Qin, D., Zhao, X.-M., Sohn, L.L., Celotta, R.J., and Whitesides, G.M.: *Adv. Mater.* **9**, 147–149 (1997)
8. Chou, S.Y.: *Science* **272**, 85–87 (1996)
9. Chou, S.Y., Krauss, P.R., and Renstrom, P.J.: *Appl. Phys. Lett.* **67**, 3114–3116 (1995)
10. Hoyer, P., Baba, N., and Masuda, H.: *Appl. Phys. Lett.* **66**, 2700–2702 (1995)
11. Masuda, H., and Fukada, K.: *Science* **268**, 1468–1466 (1995)
12. Piner, D., Zhu, J., Xu, F., and Hong, S., and Mirkin, C.A.: *Science* **283**, 661–63 (1999)
13. Hong, S., Zhu, J., and Mirkin, J.: *Science* **286**, 523–525 (1999)
14. Ivanisevic, A., and Mirkin, C.A.: *J. Am. Chem. Soc.* **123**, 7887–7889 (2001)

15. Houel, A., Tonneau, D., Bonnail, N., Dallaporta, H., and Safarov, V.: *J. Vac. Sci. Technol. B* **20** (6), 2337 (2002)
16. Mischovsky, N.M., and Tsong, T.T.: *Phys. Rev. B* **46**, 2640–2643 (1992)
17. Mamin, H.J., Guethner, P.H., and Rugar, D.: *Phys. Rev. Lett.* **65**, 2418–2421 (1990)
18. Masher, C., and Damaschke, B.: *J. Appl. Phys.* **75**, 5438–5440 (1994)
19. Hsiao, G.S., Penner, R.M., and Kingsley, J.: *Appl. Phys. Lett.* **64**, 1350–1352 (1994)
20. McCord, M.A., Kern, D.P., and Chang, T.H.P.: *J. Vac. Sci. Technol. B* **6**, 1877–1880 (1988)
21. Rauscher, H., Behrendt, F., and Behm, R.J.: *J. Vac. Sci. Technol. B* **15**, 1373–1377 (1997)
22. Marchi, F., Bouchiat, V., Dallaporta, H., Safarov, V., and Tonneau, D.: *J. Vac. Sci. Technol. B* **18**, 1171–1176 (2000)
23. Legrand B., Deresmes, D., and Stievenard, D.: *J. Vac. Sci. Technol. B* **20**, 862–870 (2002)
24. Crommie, M.F., Lutz, C.P., and Eigler, D.M.: *Science* **262**, 218–220 (1993)
25. Fischlock, T.W., Oral, A., Egdell, R.G., and Pethica, J.B.: *Nature* **404**, 743–745 (2000)
26. Eigler, D.M., and Scheitwer, E.K.: *Nature* **344**, 524–526 (1990)
27. Kirakosian, A., McChesney, J.L., Bennwitz, R., Crain, J.N., Lin, L., and Himpfel, F.J.: *Surf. Sci.* **498**, 109–112 (2002)
28. Baski, A.A., Erwin, S.C., Turner, M.S., Jones, K.M., Dickinson, J.W., and Carisle, J.A.: *Surf. Sci.* **476**, 22–34 (2001)
29. Erwin, S., and Weitering, H.H.: *Phys. Rev. B* **81**, 2296–2299 (1998)
30. Palmino, F., Ehret, E., Louay, L., Labrune, J.C., Lee, G., Hanchul, K., and Themlin, J.M.: *Phys. Rev. B* **67**, 195413 (2003)
31. Soukiassian, P., Semond, F., Mayne, A., and Dujardin, G.: *Phys. Rev. Lett.* **79**, 2498 (1997)
32. Staub, R., Toerker, M., Fritz, T., Schmitz-Hübsch, T., Sellam, F., and Leo, K.: *Langmuir* **14**, 6693 (1998)
33. Zach, M.: *Science* **290**, 2120–2123 (2000)
34. Munford, M.L., Maroun, F., Cortes, R., Allongue, P., and Pasa, A.A.: *Surf. Sci.* **537**, 95–112 (2003)
35. Gambardella, P., Blanc, M., Brune, H., Kuhnke, K., and Kern, K.: *Phys. Rev. B* **61**, 2254–2262 (2000)
36. Ahn, J.R., Kim, Y.J., Lee, H.S., Hwang, C.C., Kim, B.S., and Yeom, H.W.: *Phys. Rev. B* **66**, 153403
37. Ragan, R., Ahn, C.C., and Atwater, H.A.: *Appl. Phys. Lett.* **82**, 3439–3441 (2003)
38. Preinesberger, C., Becker, S.K., Vandr e, S., Kalka, T., and D ahne, M.: *J. Appl. Phys.* **91**, 1695–1697 (2002)
39. Nogami, J., Liu, B.Z., Kartov, M.V., and Ohbuchi, C.: *Phys. Rev. B* **63**, 233305 (2001)
40. Duan, X.F., and Lieber, C.M.: *J. Am. Chem. Soc.* **122**, 188 (2000)
41. Karaguchi, K., Katsuyama, T., Kiruma, K., and Ogawa, K.: *Appl. Phys. Lett.* **60**, 745 (1992)
42. Xu, D., Xu, Y., Chen, D., Guo, G., Gui, L., and Tang, Y.: *Adv. Mater.* **12**, 520 (2000)



43. Sander, S., Gronsky, R., Sands, T., and Stacy, A.M.: *Chem. Mater.* **15**, 335–339 (2003)
44. Park, W.I., Kim, D.H., Jung, S.W., and Yi, G.C.: *Appl. Phys. Lett.* **80**, 22 (2002)
45. Gimzewski, J.K., and Möller, R.: *Physica B* **36**, 1284–1287 (1986)
46. van Wees, B.J., van Houten, H.H., Beenakker, C.W.J., Williamson, J.G., Kouwenhoven, L.P., van der Marel, D., and Foxon, C.T.: *Phys. Rev. Lett.* **60**, 848–850 (1988)
47. Wharam, D.A., Thornton, T.J., Newbury, R., Pepper, M., Hamed, H., Frost, J.E.F., Hasko, D.G., Peacock, D.C., Ritchie, D.A., and Jones, G.A.C.: *J. Phys. C* **21**, L209–L214 (1988)
48. Muller, C.J., van Ruitenbeek, J.M., and de Jongh, L.J.: *Physica C* **191**, 485–504 (1992)
49. Datta, S.: *Electronic Transport in Mesoscopic Systems*, Cambridge University Press, Cambridge (1997)
50. Imry, Y.: *Introduction to Mesoscopic Physics*, Oxford University Press, Oxford (1997)
51. Landauer, R.: *IBM J. Res. Dev.* **1**, 223–231 (1957)
52. Landauer, R.: *Phil. Mag.* **21**, 863–867 (1970)
53. Landauer, R.: *IBM J. Res. Dev.* **32**, 306 (1988)
54. Büttiker, M.: In *Electronic Properties of Multilayers and Low Dimensional Semiconductors*, Plenum Press, New York (1990) pp. 51–73
55. Büttiker, M.: *Phys. Rev. B* **46**, 12485–12507 (1992)
56. Scheer, E., Agraït, N., Cuevas, J.C., Levy Yeyati, A., Rudolph, B., Martin-Rodero, A., Rubio Bollinger, G., van Ruitenbeek, J.M., and Urbina, C.: *Nature* **394**, 154–157 (1998)
57. Mahan, J.: *Many-Particle Physics*, Plenum Press, New York (1990)
58. Agraït, N., Yeyati, A.L., and van Ruitenbeek, J.M.: *Phys. Rep.* **377**, 81–279 (2003)
59. Brandbyge, M., Taylor, J., Stokbro, K., Mozos, J.L., and Ordejon, P.: *Phys. Rev. B* **65**, 165401 (2002)
60. Palacios, J.J., Perez-Jimenez, A.J., Louis, E., SanFabian, E., and Verges, J.A.: *Phys. Rev. B* **66**, 035322 (2002)
61. Todorov, T.N.: *J. Phys.: Condens. Matter* **12**, 8995–9006 (2000)
62. van Houten, H., Beenakker, C.W.J., and Staring, A.A.M.: Coulomb-blockade oscillations in semiconductor nanostructures. In: H. Grabert and M.H. Devoret (Eds.), *Single Electron Tunneling*, Plenum Press, New York (1991) pp. 167–216
63. Dong, Z.C., Fujita, D., and Nejh, H.: *Phys. Rev. B* **63**, 115–402 (2001)
64. Hitosugi, T., et al.: *Phys. Rev. Lett.* **82**, 4034 (1999)
65. Ohnishi, H., Kondo, Y., and Takayanagi, K.: *Nature* **395**, 780 (1998)
66. Reed, M.A., Zhou, C., Muller, C.J., Burgin, T.P., and Tour, J.M.: *Science* **278**, 252 (1997)
67. Onipko, A.: *Phys. Rev. B* **59**, 9995 (1999)
68. Hall, L.E., Reimers, J.R., Hush, N.S., and Silverbrook, K.: *J. Chem. Phys.* **112**, 1510 (2000)
69. Magoga, M., and Joachim, C.: *Phys. Rev. B* **57**, 1820 (1998)
70. Mujica, V., Roitberg, A.E., and Ratner, M.J.: *Chem. Phys.* **112**, 6834 (2000)
71. Nakanishi, S., and Tsukada, M.: *Surf. Sci.* **438**, 305 (1999)
72. Tsukada, M., Kobayashi, N., Brandbyge, M., and Nakanishi, S.: *Prog. Surf. Sci.* **64**, 139 (2000)

73. Magoga, M., and Joachim, C.: Phys. Rev. B **59**, 16011 (1999)
74. Dash, L.K., and Fisher, A.J.: J. Phys.: Condens. Matter **13**, 5035 (2001)
75. Segovia, P., Purdie, D., Hengsberger, M., and Baer, Y.: Nature **402**, 504–507 (1999)
76. Peierls, R.E.: *Quantum Theory of Solids*, Clarendon Press, Oxford (1955)
77. Himpfel, F.J., Altmann, K.N., Bennewitz, R., Crain, J.N., Kirakosian, A., Lin, J.L., and McChesney, J.L.: J. Phys.: Condens. Matter **13**, 11097–11113 (2001)
78. Ness, H., Shevlin, S.A., and Fisher, A.J.: Phys. Rev. B **63**, 125422 (2001)
79. Petrov, E.G., and Hänggi, P.: Phys. Rev. Lett. **86**, 2862 (2001)
80. Petrov, E.G., May, V., and Hänggi, P.: Chem. Phys. **281**, 211 (2002)
81. Hettler, M.H., Schoeller, H., and Wenzel, W.: Europhys. Lett. **57**, 571 (2002)
82. Hettler, M.H., Wenzel, W., Wegewijs, M.R., and Schoeller, H.: Phys. Rev. Lett. **90**, 076805 (2003)
83. Agraït, N., Untiedt, C., Rubio-Bollinger, G., and Viera, S.: Chem. Phys. **281**, 231–234 (2002)
84. Donhauser, Z.J., Mantooth, B.A., Kelly, K.F., Bumm, L.A., Monnell, J.D., Stapleton, J.J., Price Jr., D.W., Allara, D.L., Tour, J.M., and Weiss, P.S.: Science **292**, 2303 (2001)
85. Cui, Y., Wei, Q., Park, H., and Lieber, C.: Science **293**, 1289–1292 (2001)
86. Duan, X., and Lieber, C.: Adv. Mat. **12**, 298–302 (2000)

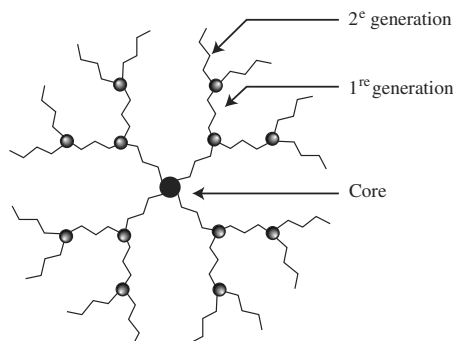
## Nano-Objects

J.-F. Nierengarten, J.-L. Gallani, and N. Solladié

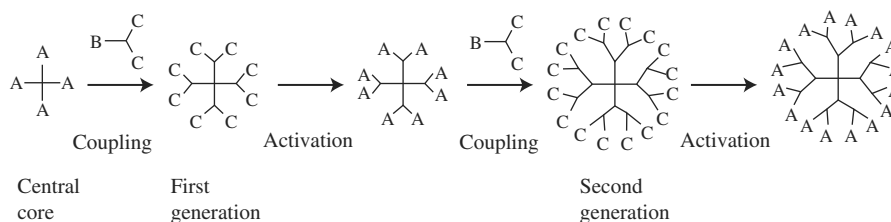
For the physicist, nanoscience is the business of making things ever smaller (the top-down approach), whereas the chemist adopts the opposite attitude, starting with atoms or molecules and building up to objects of nanometric size, designing in specific properties on the way (the bottom-up approach). This chapter discusses the various strategies that chemists have devised to build giant molecular objects. To begin with, we shall describe a family of molecules with branched or tree-like structure, known as dendrimers [1, 2]. These compounds, which are sometimes as big as natural proteins, are fabricated using the ideas of ‘standard’ molecular chemistry, i.e., the chemistry of covalent bonding. A considerable effort is often required to synthesise such objects, and an alternative approach consists in associating several simple chemical species that are easier to synthesise, by exploiting intermolecular forces. This is supramolecular chemistry [3]. These non-covalent interactions can be used to form nano-objects with characteristics as well defined as those of standard molecular compounds. They can also provide several examples of organised molecular assemblies, thus opening up prospects for the design of molecular machines. Non-covalent interactions can also lead to polymolecular assemblies resulting from the spontaneous association of an undetermined number of components into a phase exhibiting different degrees of nanoscopic organisation, such as layers, membranes, vesicles and micelles. We shall end the chapter with a description of these objects.

### 10.1 Dendrimers

A dendrimer is a macromolecule made up of basic units called monomers, which aggregate into a branched structure around a multifunctional core (see Fig. 10.1). This structure builds up by iteration of a sequence of reactions. At the end of each reaction cycle, a new generation is obtained, and there is an increasing number of identical branches at each stage. In contrast with polymers, whose synthesis leads to a distribution of structures, dendrimers are



**Fig. 10.1.** Schematic representation of a second generation dendrimer



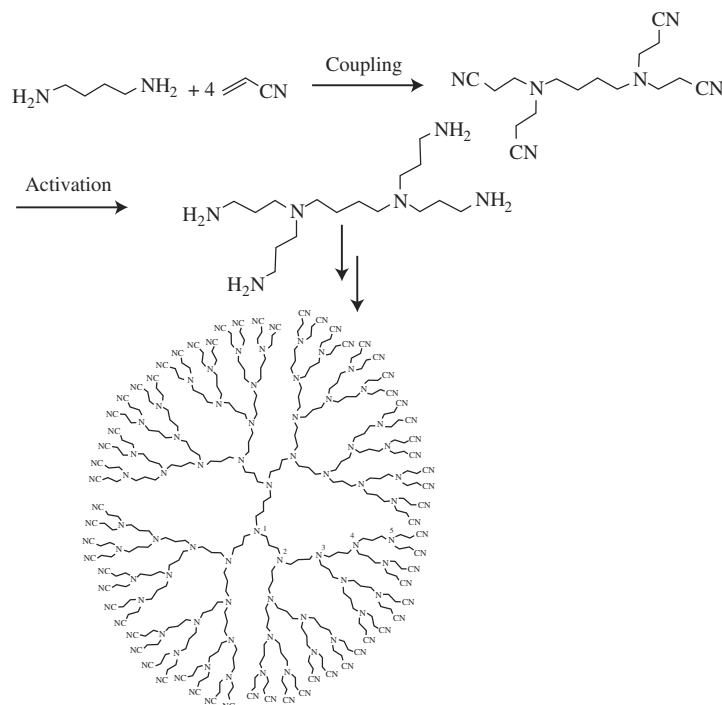
**Fig. 10.2.** Synthesis of dendrimers by the divergent method

synthesised using methods which in principle control their structure at each stage in the construction. Due to their novel branched structure, dendrimers have special properties. It would be impossible to summarise here all the work done on these objects and the reader is referred to [1,2]. We shall limit the discussion to describing the two main strategies for synthesising these compounds.

### 10.1.1 Divergent Synthesis

The divergent approach was used by Vögtle, Tomalia and Newkome to prepare the first dendrimers [1,2]. Dendrimer growth is achieved from a central unit by implementing a cycle of addition reactions followed by activation. This cycle begins when functional groups placed on the periphery of the central unit couple with the focal group of the monomer unit (see the coupling step in Fig. 10.2). This leads to the formation of a first generation dendrimer. The new periphery is inactive with respect to the monomer.

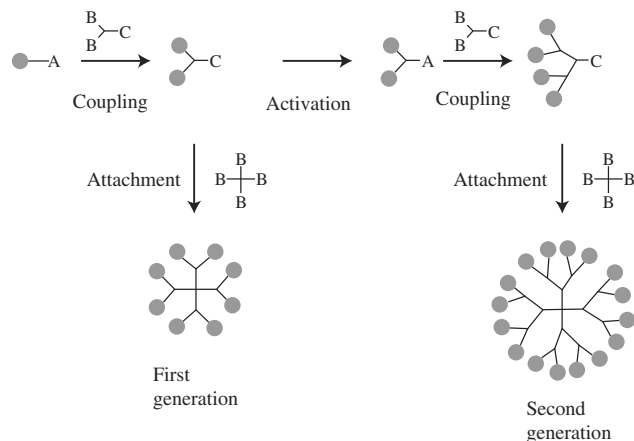
The formation of a second generation thus requires the activation of all the functional groups situated on the periphery (see the activation step in Fig. 10.2). This means converting them into active species capable of reacting with a further monomer unit. A new coupling step then leads to the formation of a second generation dendrimer. Repeating these activation and coupling steps leads to rapid growth of a dendritic structure. Although this



**Fig. 10.3.** An example of dendrimer preparation using the divergence strategy

approach may appear to be direct and easy to use from a conceptual viewpoint, difficulties sometimes arise in the synthesis. One obstacle lies in the increasing number of functional groups on the surface as the dendrimer grows bigger. Consequently, to obtain high generation compounds, a large number of reactions has to occur on the same molecule and this number rises exponentially with the generation. Hence, a total functionalisation of the surface is sometimes difficult to achieve for high generation dendrimers. Another difficulty arising from this situation lies in the problem of purification. It is often a delicate matter to separate the required dendrimer from others with functional groups on the surface that have not reacted, especially if the masses, sizes or properties of these products are very similar.

To illustrate this strategy, Fig. 10.3 gives a schematic view of the preparation of dendrimers of poly(propylene imine) type. The coupling step consists in getting each primary amine function of the core to react with two equivalents of acrylonitrile. The reduction of the nitrile functions located on the periphery is then used to generate amine functions, whereupon a new coupling step can occur, leading to a higher generation dendrimer. By repeating these successive coupling and reduction steps, it is thus possible to prepare the fifth generation dendrimer with 64 peripheral nitrile functions. This molecular



**Fig. 10.4.** Dendrimer synthesis by the convergent method

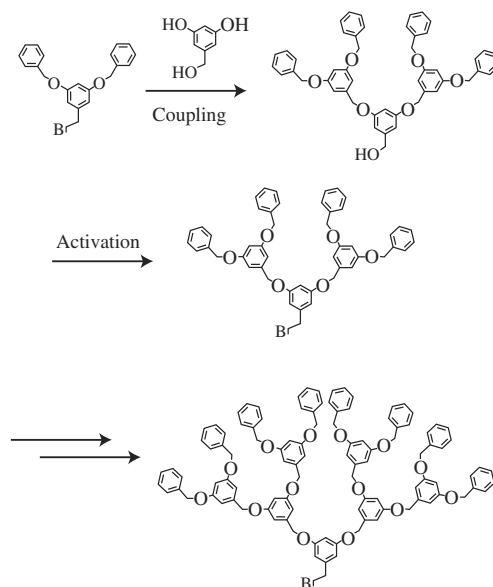
object has a more or less spherical shape with diameter of the order of 2.6–3.1 nm and volume about 17.5 nm<sup>3</sup>.

### 10.1.2 Convergent Synthesis

The second approach, known as convergent synthesis, was developed by Hawker and Fréchet [1,2]. The first step is to construct the dendritic branches, called dendrons (see Fig. 10.4). These branches are made by an iterative sequence of coupling and activation reactions. A first dendron is prepared from a monomer and a basic unit (activation step in Fig. 10.4). The next generation dendron is built by coupling this first activated dendron to each of the functional groups of a further monomer unit (coupling step in Fig. 10.4). Hence, by iterating this activation/coupling process, higher generation dendrons can be made. In the final stage, the dendrons are attached to a central core to produce the dendrimer.

One of the advantages of this method is that not many reactions are required per molecule at each stage (2 or 3 suffice). The number of secondary products is therefore limited. Moreover, defective products have very different masses from the final dendrimer and it becomes much easier to purify the result. However, it should be noted that one of the difficulties with this synthesis arises due to the central position of the reactive groups. Indeed, for high generations, the focal point is more and more isolated by the presence of the branches, which tend to reduce its reactivity. A reduction and slowing down of the reactivity of the branches are thus commonly observed during their growth.

The convergent method would appear to be the best for synthesising monodispersed dendrimers, but the problems of steric hindrance still prohibit the preparation of high generation products. On the other hand, divergent



**Fig. 10.5.** Example of the synthesis of dendritic branches

synthesis is the most appropriate for synthesising high generation dendrimers, even though it is difficult to avoid the appearance of defects in these structures. In this sense, the convergent and divergent methods can be said to be complementary.

To illustrate the convergent strategy, Fig. 10.5 shows the preparation of dendrons of type poly(benzyl ether). The transition from one generation to another is made in two steps: first, there is an activation step in which a benzyl alcohol function is transformed into a benzyl bromide function; then there follows a coupling stage in which this bromide couples with 3,5-dihydroxybenzyl alcohol. The dendrons formed in this way can subsequently be attached to a central core to form the dendrimer.

## 10.2 Supramolecules

Today the importance of weak non-covalent interactions in biology is widely accepted. One may think, for example, of the human gene pool, made up of deoxyribose nucleic acids (DNA). These exist in our cells in the form of a double helix that is stabilised by non-covalent interactions called hydrogen bonds. Two strands of DNA join together in an antiparallel manner to form a double helix with the help of weak bonds which can be broken temporarily to allow transcription of the DNA into an RNA (ribose nucleic acid) messenger, thereby allowing the synthesis of specific proteins. We may thus conclude that

the perpetuation of our gene pool rests upon the possibility of breaking and rebuilding a whole series of weak non-covalent bonds.

What is striking about natural phenomena is the complexity of the systems involved, devoted to highly specific tasks. Supramolecular chemistry is concerned with assemblies of several molecules into non-covalent constructions, in the way illustrated by biological systems. The problem here is one of molecular recognition, i.e., a complementarity of shape, size and chemical functions which exploits the weak intermolecular interactions that may exist over short distances between several molecules. In decreasing order of interaction energy, the non-covalent forces are: complexation forces due to metal cations, hydrogen bonds, hydrophobic interactions,  $\pi$  interactions, and charge transfer interactions.

### 10.2.1 Self-Assembly by 3D Template Effect Induced by a Metal Cation

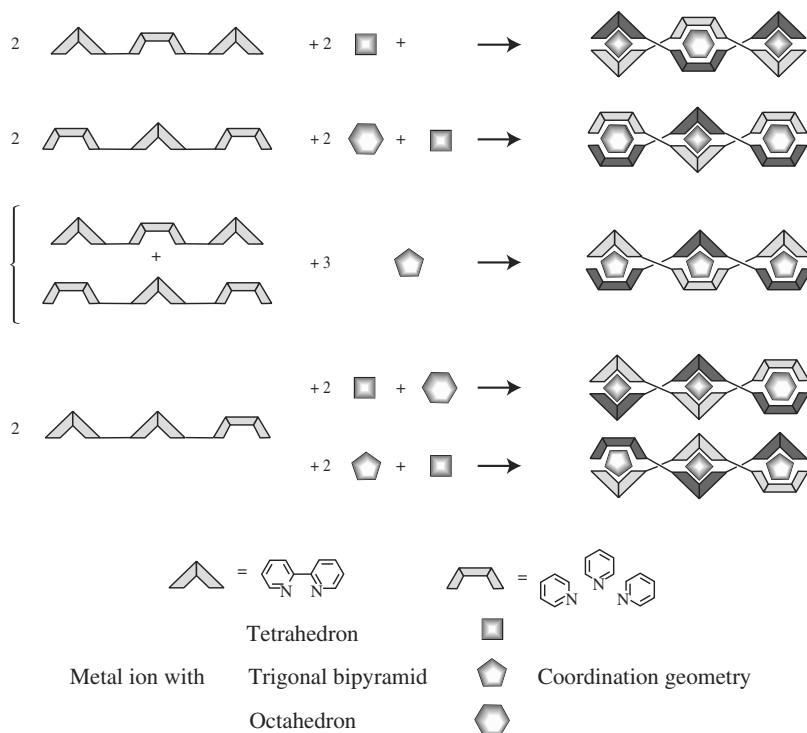
Metal–ligand bonds are strong enough and varied enough from a structural point of view to allow the construction of molecular architectures containing several subunits. Although this form of recognition bears little resemblance to the kind of self-assembly observed in biological systems, the study of molecular architectures self-assembled by metal–ligand bonds is well developed. The reason is that this type of non-covalent bond is strong enough for some such constructions to survive in an aqueous medium. Hence, by exploiting the template effect, it has been possible to consider building more sophisticated molecular architectures than those that could be designed using only the chemistry of molecular synthesis.

Let us consider some examples which illustrate the potential of these metal–ligand bonds and the template effect. A special tribute should be made in this context to J.-M. Lehn who was awarded the Nobel Prize in 1987 for his work on supramolecular chemistry. As an example, in one of his publications, he describes the self-assembly of several double helices using the template effect induced by different metals accepting different coordination geometries on ‘tritopic’ molecular strands, i.e., capable of recognising and complexing three metal cations, carrying different bi- or tridentate motifs (see Fig. 10.6) [4].

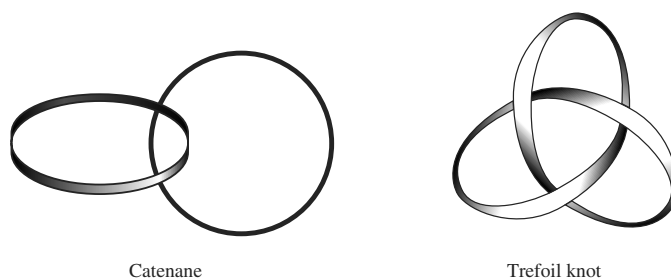
This technique allows the choice of assembling a single type of strand, or several. One can also choose the way in which the molecular strands are matched, simply by changing the metal cations involved in the recognition process. The 3D template effect induced by Cu(I) cations can also be used to assemble several precursors so as to establish directed covalent bonds for subsequent synthesis of molecules exhibiting new types of isomerism and with non-trivial topology (see Fig. 10.7).

For example, the formation of a Cu(I) bis-phenanthroline complex can be used to assemble a pre-catenate type of structure which is stable enough to allow a double cyclisation reaction and to obtain a catenate, i.e., a complex of



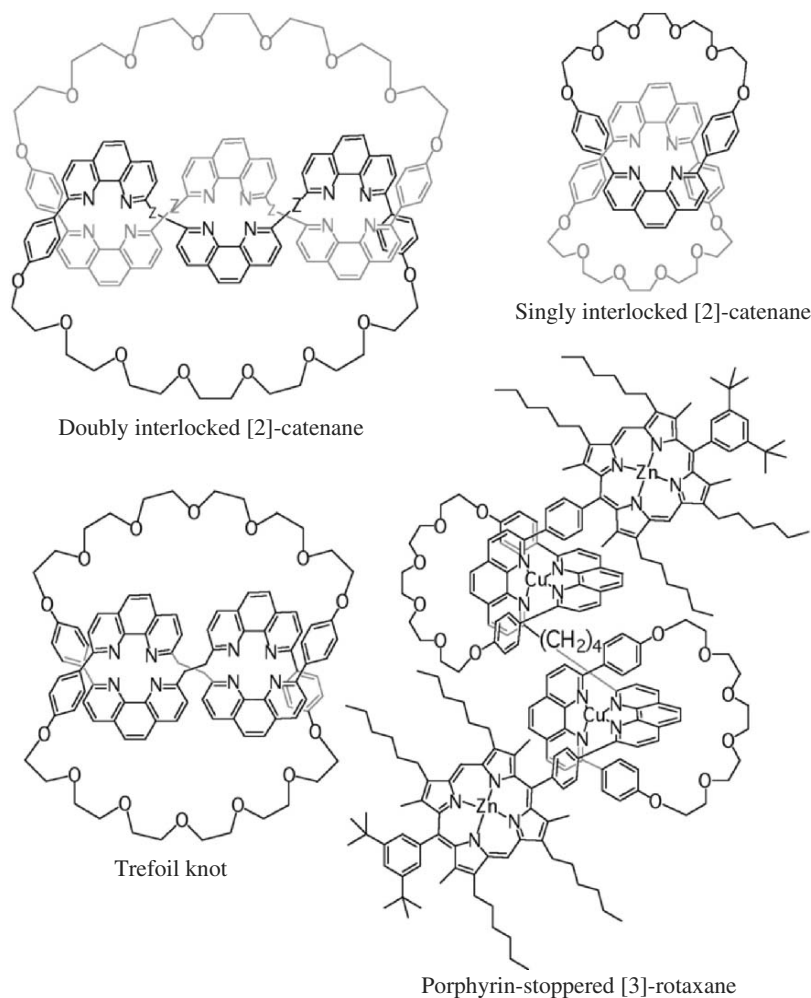


**Fig. 10.6.** Schematic representation of the self-assembly of linear tritopic ligands (*left*) containing a given sequence of complexing units destined for certain well defined metal cations (*center*) to yield, in the expected and predicted way, double helix complexes called helicates (*right*) [4]



**Fig. 10.7.** Example of two molecules with non-trivial topology. Catenane is an isomer with two macrocycles, while the trefoil knot comprises a single macrocycle

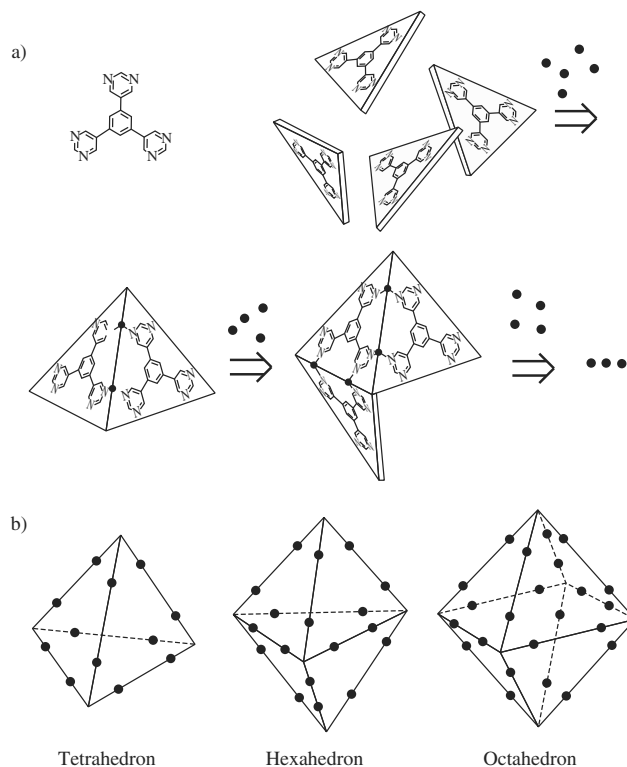
Cu(I) with two interlocked macrocycles. The molecule obtained after decomplexation is called catenane (see Fig. 10.7) [5]. Such a molecule is said to have non-trivial topology because it cannot be drawn on a sheet of paper without crossing itself, i.e., without lifting the pencil. Indeed, the two macrocycles



**Fig. 10.8.** Examples of two catenanes, a trefoil knot and a rotaxane, synthesised by J.P. Sauvage and coworkers [5–7]

remain bound together after decomplexation and can only be separated by cutting one of them.

The 3D template effect induced by Cu(I) is also used to synthesise rotaxanes (see Fig. 10.8) [6]. The template effect is used here to thread one or more macrocycles onto a molecular string. The presence of blocking groups at each end of the strand after demetalation of the complexes prevents the macrocycles from dethreading. Finally, the formation of a double helix between two multi-phenanthroline strands self-assembled around several Cu(I) cations was used to synthesise the first molecular knots and doubly interlocked catenanes (see Fig. 10.8) [7].

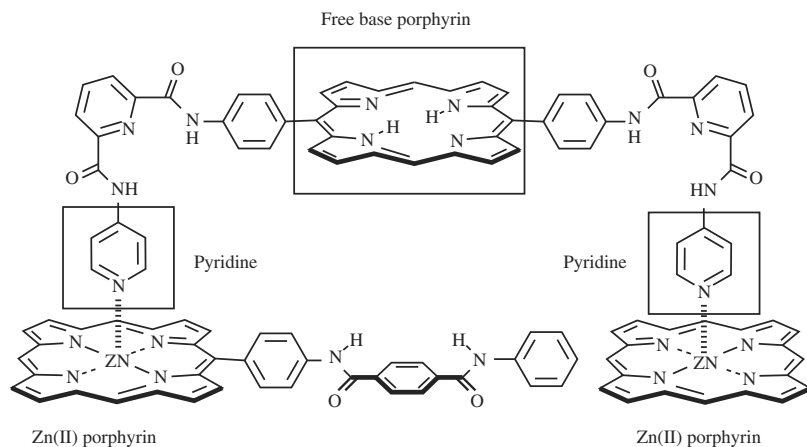


**Fig. 10.9.** Self-assembly between ligand 1 and metal cations (a) and representations of the resulting polyhedral architecture (b). Courtesy of M. Fujita et al. [8]

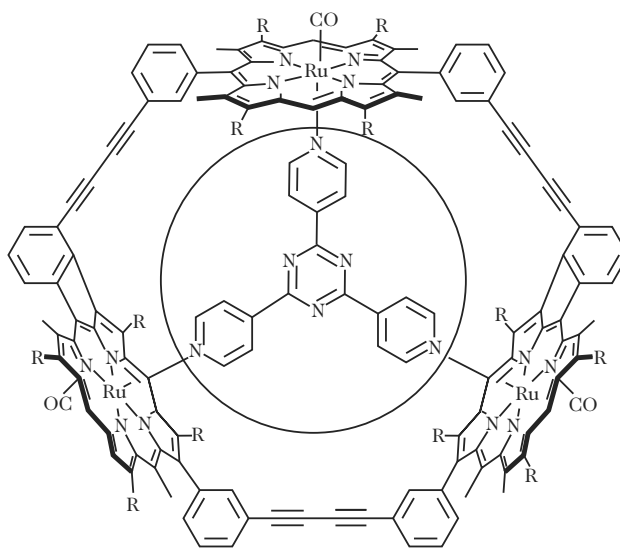
Nanoscale capsules can also be synthesised with the help of the template effect, using coordination chemistry. These are formed by self-assembly of 24 components, including 18 metal cations and 6 triangular hexadentate ligands (see Fig. 10.9) [8]. The nanocages made in this way comprise 6 triangles arranged into a hexahedron and enclosing a volume of  $900 \text{ \AA}^3$ .

The ability of metalated porphyrins to accept axial coordinations has also been used in supramolecular chemistry to synthesise a self-assembled macrocycle by axial complexation of two Zn(II) porphyrins of one strand by the two pyridine groups substituting the free base porphyrin of the other strand. This macrocycle comprises two strands that are complementary both in function and in geometry, in such a way as to allow the self-association of the pyridines on the Zn(II) porphyrin by axial complexation (see Fig. 10.10) [9].

By exploiting the axial coordination of porphyrins, it has also been possible to carry out self-assembly between a macrocycle composed of three Zn(II) porphyrins and a tris-pyridine with size and shape complementary to its cyclic host (see Fig. 10.11) [10].

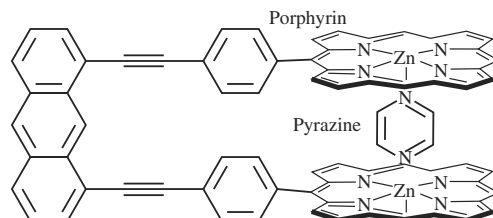


**Fig. 10.10.** Macrocyclic structure formed by axial complexation of two Zn(II) porphyrins (*lower strand*) by the two pyridine groups substituting one central free base porphyrin (*upper strand*). Courtesy of C. Hunter et al. [9]

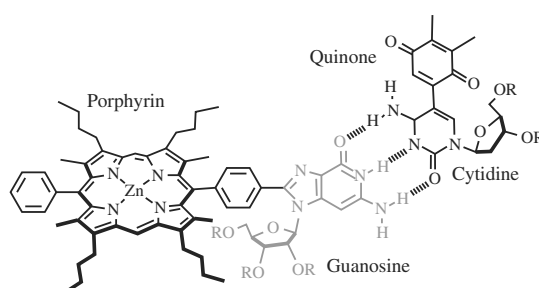


**Fig. 10.11.** Molecular recognition between cyclic tris-porphyrin and tris-pyridine (*centre*) with complementary size and shape. A beautiful example of complementarity between two self-assembled partners. Courtesy of J. Sanders et al. [10]

Apart from the purely structural interest of this kind of construction, recognition by axial coordination can be exploited to modify the physico-chemical properties of a molecule. Figure 10.12 shows the synthesis of a co-facial bis-porphyrin tweezer and the formation of a guest/host complex with a bidentate base, namely, pyrazine [11]. By inserting this base in a suitably



**Fig. 10.12.** Insertion of a pyrazine guest into the cavity of a cofacial bis-porphyrin tweezer. Generation of electronic coupling between the two porphyrins by formation of a guest/host complex. Courtesy of N. Solladié et al. [11]

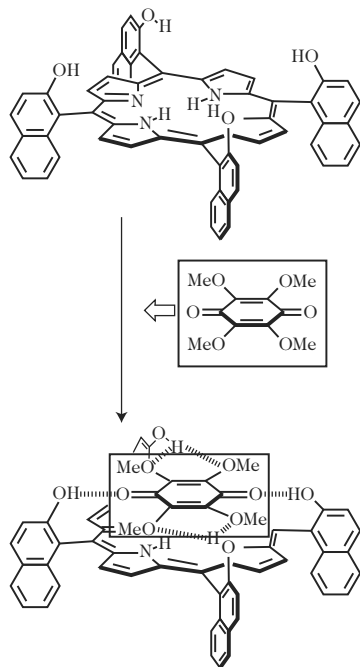


**Fig. 10.13.** Self-assembly between a guanosine entity substituted by a porphyrin and a derivative of cytidine functionalised by a quinone. Courtesy of J. Sessler et al. [13]

dimensioned bis-porphyrin cavity, the physicochemical properties of this molecular tweezer are modified by generating an electronic coupling between the porphyrins.

### 10.2.2 Self-Assembly by Hydrogen Bonding

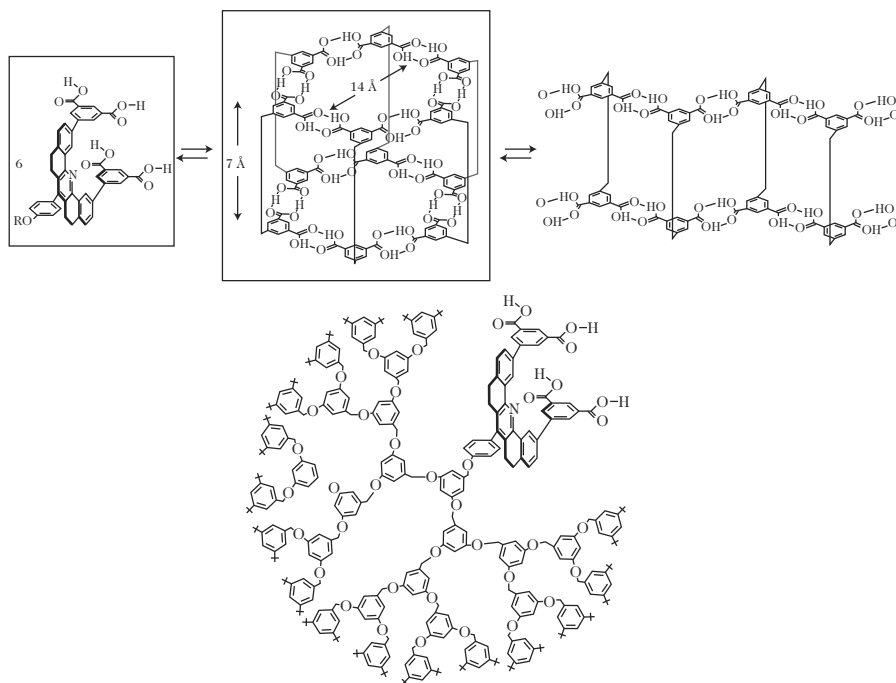
Hydrogen bonds provide perhaps the most convenient intermolecular self-assembly force by virtue of their specificity and directionality. They represent an energy of 5–10 kcal/mol. Among those systems bound by hydrogen bonds, two-dimensional structures are by far the most widespread in supramolecular chemistry [12]. For the main part, these systems use rigid, plane heterocycles as assemblers. For example, consider the dimer shown in Fig. 10.13, which comprises a guanosine entity substituted by a porphyrin and a derivative of cytidine functionalised by a quinone [13]. The nucleic base guanine is a heterocycle that can establish three hydrogen bonds with its complementary partner, viz., the nucleic base cytosine. Guanine carries only two of the three hydrogen atoms involved in the assembly and is said to be the donor of two hydrogen bonds and the acceptor of one. Conversely, cytosine is the acceptor of two hydrogen bonds and donor of one.



**Fig. 10.14.** *meso-α, α, α, α-tetra(2-hydroxy-1-naphthyl)-porphyrin* (*top molecule*) has been produced as a molecular host that can accept a guest tetramethoxy-*p*-benzoquinone by virtue of the six hydrogen bonds that can be established between the two partners. The molecular recognition exemplified here is used to observe electron transfer from the porphyrin to its benzoquinone guest, and plays an essential role in conditioning the physicochemical properties of the self-assembled dimer. Courtesy of T. Hayashi, H. Ogoshi et al. [14]

It is interesting to note that what we have here is not just a self-assembly phenomenon, but also a molecular recognition process. Indeed the guanine motif is perfectly matched to the cytosine motif, both by the complementarity of the atoms capable of donating or accepting hydrogen bonds in the two partners, and also by the spatial orientation of the atoms involved in the assembly. The rigidity of the heterocycles imposes an angle of  $120^\circ$  between the three atoms involved in each hydrogen bond. Supramolecular chemistry makes use of these hydrogen bonds to build ever more sophisticated systems with regard to the design of acceptors or extended two- and three-dimensional networks.

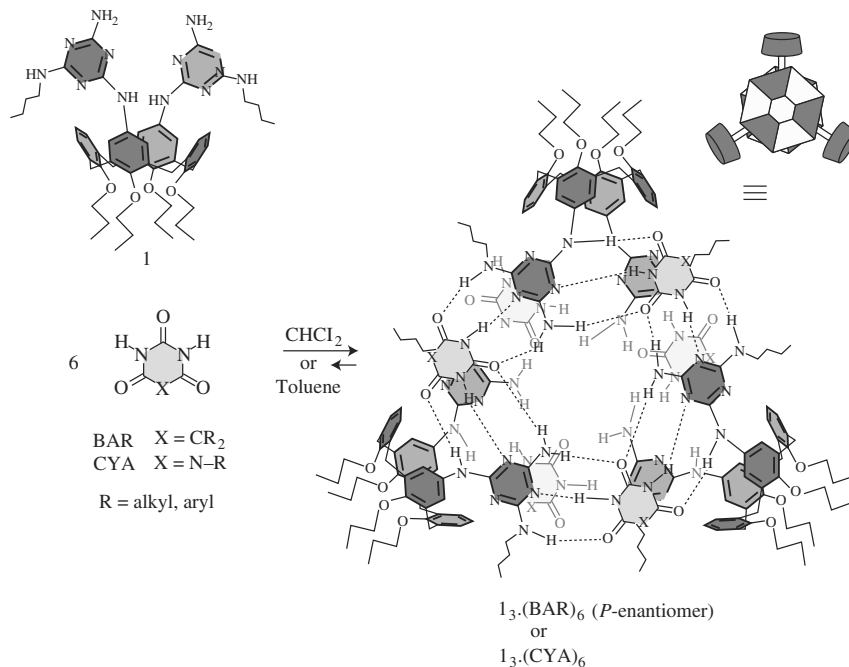
In particular, hydrogen bonds are used to synthesise potential models for natural systems. One aim is to obtain a better understanding of electron transfer processes taking place in the reaction center of photosynthesising systems, key phenomena in the conversion of solar energy into chemical energy that can be used by the cells of photosynthesising organisms and plants. For this purpose, *meso-α, α, α, α-tetra(2-hydroxy-1-naphthyl)-porphyrin* has been produced as a molecular host able to accept a guest tetramethoxy-*p*-benzoquinone by virtue of the six hydrogen bonds that can be established between the two partners (see Fig. 10.14) [14]. The molecular recognition described here is necessary for the observation of electron transfer from the porphyrin to its guest benzoquinone. It thus plays a key role in conditioning the physicochemical properties of the self-assembled dimer.



**Fig. 10.15.** Schematic representation of two molecular architectures one cyclic and one linear, created by self-assembly of 6 molecules, each of which carries four carboxylic acid groups (*top left*). *Top centre*: Schematic representation of a nanocage favoured by steric hindrance generated by dendritic chains located on each tetra-acid and illustrated in the *lower diagram*. Courtesy of S. Zimmerman et al. [15]

Like the metal–ligand bonds described earlier, hydrogen bonds have also been used to create more complex molecular architectures than can be obtained by conventional molecular chemistry. Among the most striking examples, it is worth mentioning the case of calix[4]arene, self-assembled by a network of hydrogen bonds and capable of engaging small molecules by inclusion. Another example is the spectacular self-assembly of molecular nanocages via recognition of six molecules, each of which incorporates four carboxylic acid functions (see Fig. 10.15) [15].

This assembly of hexameric cyclic compounds or cages is favoured over the formation of linear oligomers by exploiting the steric hindrance generated by the dendritic branches located on each tetra-acid entity. An even greater challenge for supramolecular synthesis is illustrated by the assembly of non-covalent hosts via 36 hydrogen bonds of three calix[4]arenes, diametrically functionalised by two melamine groups (labelled 1 in Fig. 10.16) and six N-substituted cyanuric acid molecules (CYA) [16]. These cages are made from two cyclic platforms which are themselves made from a network of hydrogen



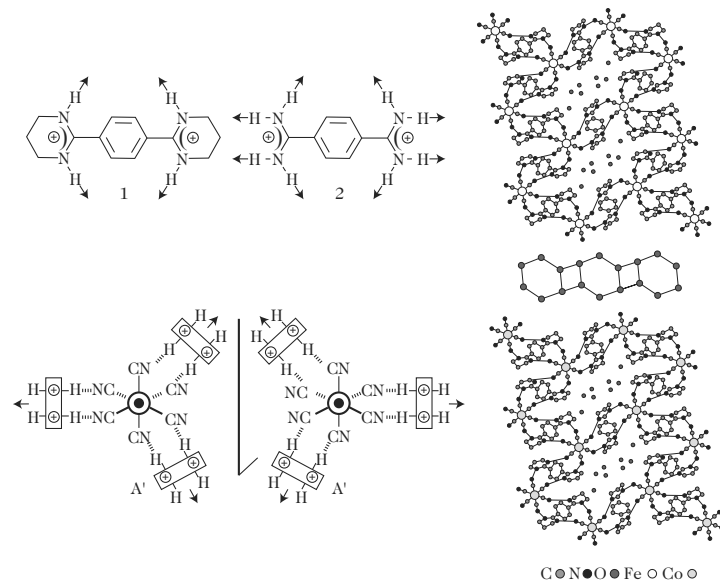
**Fig. 10.16.** Synthesis by self-assembly of double rosettes  $13 \cdot (\text{DEB})_6$  and  $13 \cdot (\text{CYA})_6$ . Molecular recognition achieved via 36 hydrogen bonds. Courtesy of P. Timmerman et al. [16]

bonds and which form the bottom and top of the molecular box, with three calixarenes for walls.

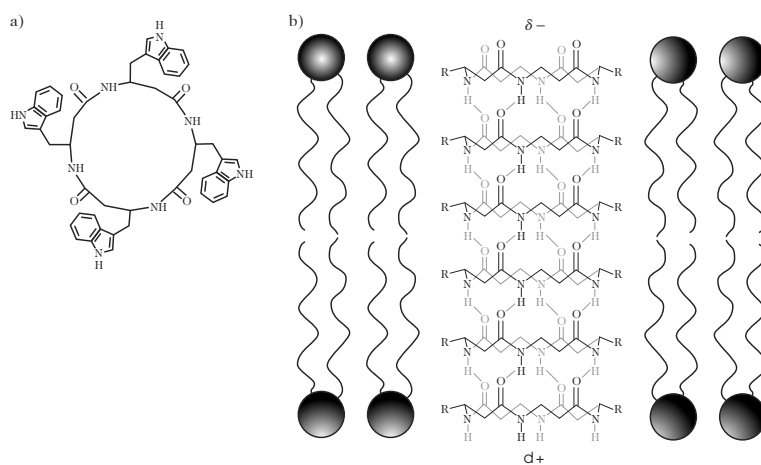
Hydrogen bonds can also be used for self-assembly of molecules into 2D networks which may one day constitute new materials with interesting properties. The subject of molecular tectonics is based on recognition between different molecules known as tectons, programmed or designed to be complementary. The idea is to produce molecular networks with preordained structures by self-assembly. As an example, Fig. 10.17 illustrates the formation of solid state 2D networks by co-crystallising different amidinium dications (labelled 1 in the figure) with the trianionic complex  $\text{M}(\text{CN})_6^{3-}$ , where  $\text{M} = \text{Pd}, \text{Pt}, \text{Mn}, \text{Cd}$  [17].

Finally, we should also mention research on synthesis of organic nanotubes, a subject attracting a great deal of interest these days. Non-covalent interactions find an ideal application here. As an example, Fig. 10.18 shows the production of ionic channels by self-assembly of cyclic  $\beta$ -tetrapeptides [18].





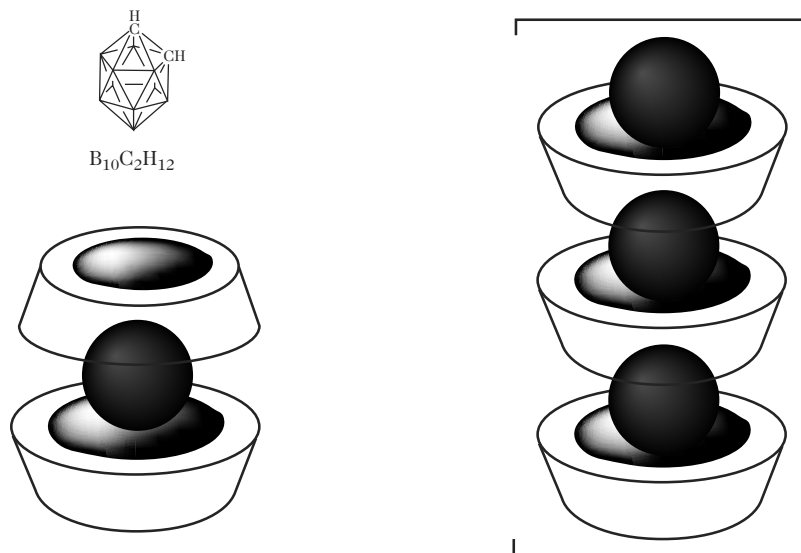
**Fig. 10.17.** Molecular tectonics. Formation of a 2D array, self-assembled via the formation of hydrogen bonds between the dicationic tecton amidinium (labelled 1) and the trianion  $M(CN)_6^{3-}$ . Courtesy of W. Hosseini et al. [17]



**Fig. 10.18.** Schematic illustration of a  $\beta^3$ -tetrapeptide (a), designed to form ionic channels through membranes (b). Courtesy of M. Ghadiri et al. [18]

### 10.2.3 Self-Assembly by Hydrophobic Interactions, $\pi$ -Interactions and Charge Transfer Interactions

Hydrophobic interactions are often used for the assembly of supramolecules. For example, the natural concavity of the cyclodextrins and cyclophanes can

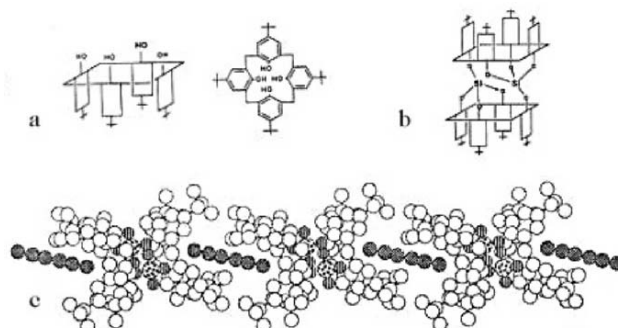


**Fig. 10.19.** Structures proposed for the formation of a 2:1 complex between a carborane (*top left*) and two  $\alpha$ -cyclodextrins (*left*) and the infinite structure of a 1:1 complex between this same carborane and  $\alpha$ -cyclodextrins (*right*). Variations in the structure are conditioned by the imposed stoichiometry. Courtesy of A. Harada, S. Takahashi et al. [19]

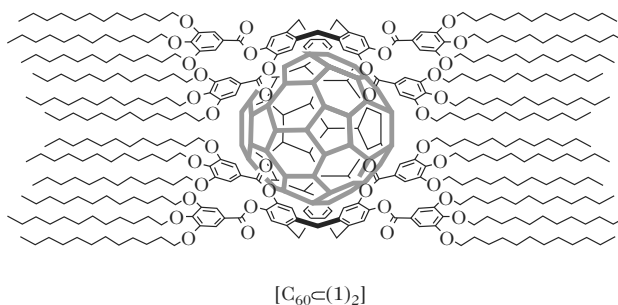
be exploited to form inclusion complexes. Indeed, the hydrophobic nature of their inner cavities can isolate apolar guest molecules from the surrounding aqueous medium. The inclusion complexes formed in the crystal state between the  $\alpha$ -,  $\beta$ - and  $\gamma$ -cyclodextrins and 1,2-dicarbododecaborane provide a specific example (see Fig. 10.19) [19].

1:1 or 1:2 complexes are obtained depending on the stoichiometry and the size of the cyclodextrin involved in the assembly process. In more direct relation to attempts to obtain new materials, Fig. 10.20 illustrates the construction in the solid state of 2D networks in which calixarenes and acetylene derivatives self-organise to form inclusion complexes [20]. This work shows that the structure of the crystal lattice can be modified by varying the length or shape of the assembling molecules. Another example is illustrated in Fig. 10.21 by the formation of a novel inclusion complex with the properties of a liquid crystal, involving two cyclotrimeratrylenes and a  $C_{60}$  molecule [21]. The two cyclotrimeratrylenes encapsulate a  $C_{60}$  to form an inclusion complex with 2:1 stoichiometry. The liquid crystal properties are due to the presence of aliphatic chains on the cyclotrimeratrylene.

The last few years have witnessed a great deal of work in the field of branched molecules or dendrimers. These adopt a similar globular conformation to proteins and, in the same way as for proteins, molecular recognition



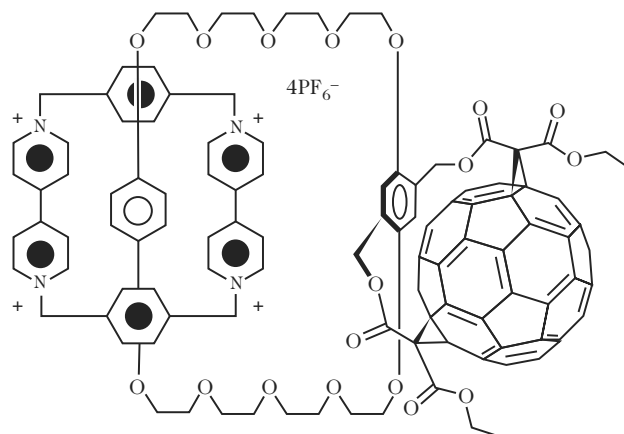
**Fig. 10.20.** Assembly of two calix[4]arenes (a) via two silicon atoms to form a koiland (b). These koilands are then assembled with the help of hexadiyne molecules (black) to form linear strings (c). Courtesy of W. Hosseini et al. [20]



**Fig. 10.21.** Formation of an inclusion complex with liquid crystal properties, involving two cyclotrimeratrylenes (black) and a C<sub>60</sub> (centre). Courtesy of J.F. Nierengarten et al. [21]

phenomena may occur, either on the periphery or in the core of the dendrimer. One thus makes the distinction here between exoreceptors and endoreceptors. As an example, it is interesting to note the synthesis of a dendrimer with 32 carboxylic acid functions on the surface, which is therefore soluble in an aqueous medium, because this shows that various apolar molecules can be solubilised in an aqueous medium by using this dendrimer [22].

Interactions between species rich in  $\pi$  electrons and electron-deficient molecules have also been exploited in supramolecular chemistry. These interactions are called charge transfer interactions, because compounds involving this type of interaction are characterised by an absorption band that can be observed by UV/visible spectroscopy and which is attributed to charge transfer from the species rich in  $\pi$  electrons to the electron-deficient species. This type of interaction, observed between  $\pi$ -donating entities such as 1,4-dialkoxybenzene or 1,5-dialkoxy-naphthalene and  $\pi$ -deficient species such as *N,N'*-dimethylpyridinium, can be exploited to assemble a whole series of



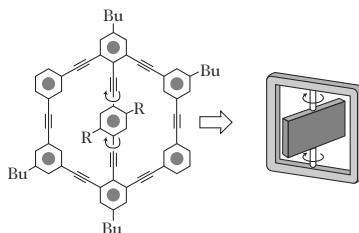
**Fig. 10.22.** [2]-catenane composed of three macrocycles of which two are interlocked. The synthesis is based upon weak interactions. Courtesy of F. Diederich, J.F. Stoddart et al. [23]

interlocking molecules. Figure 10.22 shows a [2]-catenane composed of three macrocycles, two of which are interlocked. The synthesis exploits this type of weak interaction [23].

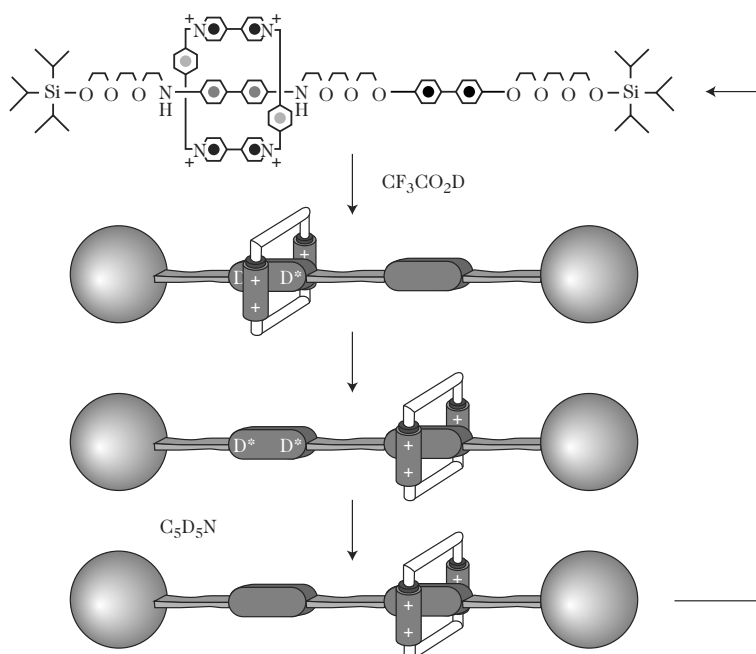
#### 10.2.4 Molecular Machines

The idea of a machine can be extended to the molecular level. A molecular machine can be defined as an assembly of a certain number of molecules which are intended to carry out mechanical movements (output) when an external stimulus (input) is applied. The term ‘molecular machine’ is only used for systems whose components undergo relatively large amplitude motions. Molecules exhibiting a simple cis/trans isomerism are not considered to be molecular machines. Likewise, systems manifesting molecular motions that are not controlled by well identified external stimuli do not fall into the category of molecular machines either. It seems crucial to extend the notion of a machine to the molecular level at a time when nanoscience and nanotechnology are developing so rapidly. J.F. Stoddart and V. Balzani have devoted a review article to this area of supramolecular chemistry [24]. A notable example is the molecular rotor, an example of which is shown diagrammatically in Fig. 10.23. It is made from a macrobicycle in which the central phenyl group is able to rotate about its axis. This movement is controlled by the size of the R groups substituting it [25].

Another type of molecular machine is the molecular shuttle. Figure 10.24 shows a [2]-rotaxane comprising an electrodeficient macrocycle and a molecular string. The latter has two different sites, both rich in  $\pi$  electrons (benzidine and biphenol) [26]. The macrocycle may locate itself around the benzidine or

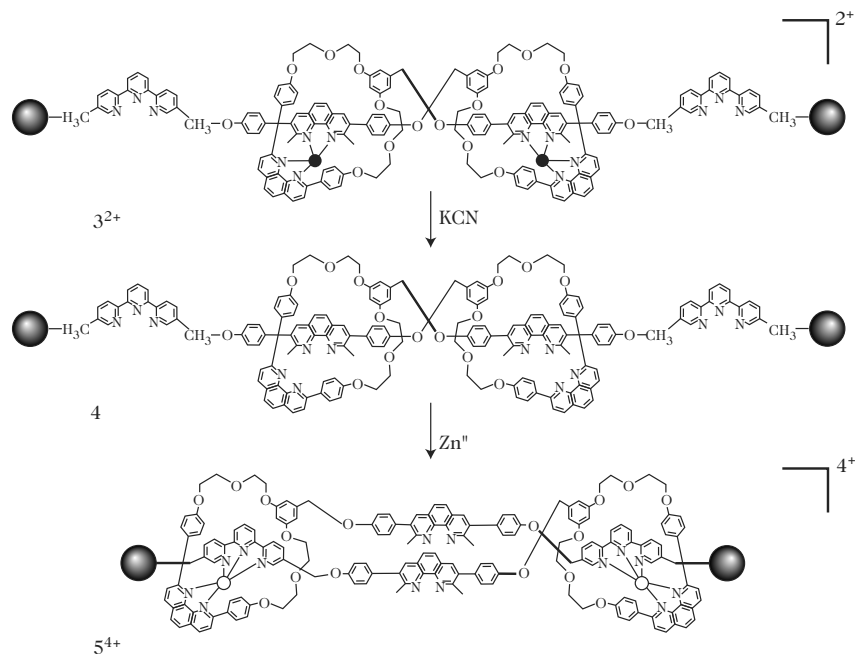


**Fig. 10.23.** Example of a molecular rotor made from a macrobicyclic whose central phenyl group is able to rotate about its axis. The motion is controlled by the size of the R groups substituting it. Courtesy of J. Moore et al. [25]



**Fig. 10.24.** Example of a molecular shuttle. A [2]-rotaxane is made from an electrodeficient macrocycle and a molecular string with two different sites, both rich in  $\pi$  electrons (benzidine in *grey* and biphenol in *black*) [24]. The macrocycle may locate itself around the benzidine entity or the biphenol entity. Courtesy of J.F. Stoddart et al. [26]

the biphenol. In solution, the two conformations are stabilised by  $\pi$  interactions or charge transfer between species rich in electrons on the strand and those poor in electrons on the macrocycle, but also by hydrogen bonds set up between the  $H\alpha$  of the bipyridinium on the macrocycle and oxygen atoms on the polyether chains of the strand. The shuttle motion of the macrocycle from one site to the other along the strand can be controlled by chemical



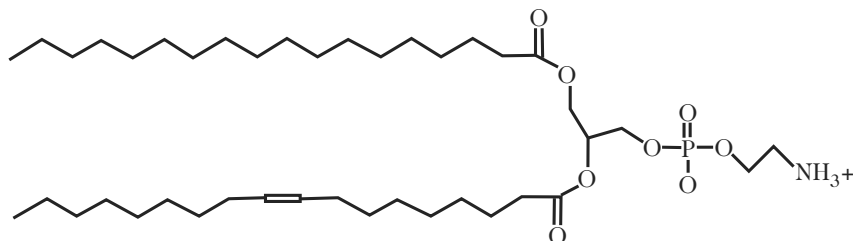
**Fig. 10.25.** Example of an artificial muscle. A molecular string made from two strands sliding one inside the other is able to contract or distend in a controlled way by altering the oxidation state of a Cu(I) or Cu(II) cation. Courtesy of J.P. Sauvage et al. [27]

or electrochemical intervention by protonating/deprotonating or oxidising/reducing the benzidine entity on the molecular string.

A third example of a molecular machine is depicted in Fig. 10.25 [27]. The subject here is an artificial muscle where a molecular string is made from two strands, one sliding within the other. This string is able to contract or distend by controlling the oxidation state of a Cu(I) or Cu(II) cation. The relative sliding motion of the two strands is permitted by the coordination geometry accepted by these two metallic cations, i.e., tetracoordination for Cu(I) and pentacoordination for Cu(II). Depending on the oxidation state of the copper, the two strands slide one inside the other to form either a tetracoordinated complex of Cu(I) with two phenanthroline ligands ( $3^{2+}$ ), or a pentacoordinated complex of Cu(II) with two different ligands, phenanthroline and terpyridine ( $5^{4+}$ ).

### 10.3 Polymolecular Assemblies

The invention of the scanning tunneling microscope (STM) has opened unbelievable prospects in the field of imaging, but it has also made it possible



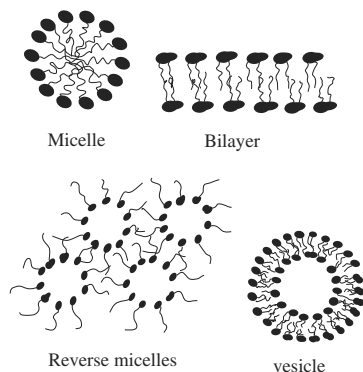
**Fig. 10.26.** Molecule of phosphatidyl ethanolamine, very common in cell membranes. The hydrocarbon chains are hydrophobic, whilst the phosphate and amine groups are hydrophilic. The double bond on one of the chains is not without significance. Its presence increases the temperature range in which the molecule is stable and improves the flexibility of biological membranes

to manipulate individual atoms and molecules [28]. The only problem when manipulating atoms or molecules by STM is that it is extremely slow. It would be quite unrealistic today to hope that it could lead to an industrial or even semi-industrial production of any kind of functional structure whatever. The most promising alternative approach is self-assembly, either spontaneous or directed. One then has the benefit of a massively parallel technique, with no obvious limiting size, the main obstacle being defects or gaps. However, one may envisage the production of highly redundant structures, since their size will never become problematic.

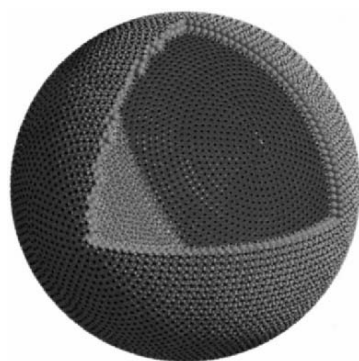
### 10.3.1 Self-Assembly in the Bulk

Historically, the first molecular assemblies appeared with the advent of life on Earth. The lipid membrane which surrounds cells (and the organelles in cells) is shared universally by all living beings, except of course for the viruses which have a protein capsid, also self-assembled. As it evolved, this membrane became more complex in order to fulfill highly elaborate functions which go far beyond the original simple role of protecting the contents of the cell from the outside world. These lipid membranes are just one special case of supramolecular architectures formed spontaneously by certain molecules called amphiphiles. Such molecules have the particular feature that one part of them, known as the polar head, is polar and hydrophilic, whilst the other is hydrophobic, often called the hydrophobic tail. The molecule in Fig. 10.26 serves as an example.

In solution, these molecules spontaneously form supramolecular assemblies with structures that depend on a great many parameters, such as the relative size of the hydrophilic and hydrophobic parts, the nature and temperature of the solvent, the possible presence of dissolved salts, and the pH. Several typical structures are illustrated in Fig. 10.27. Micelles and membranes have dimensions of the order of one or two molecular lengths, i.e., a few nanometers.



**Fig. 10.27.** Typical structures obtained by self-assembly of amphiphilic molecules

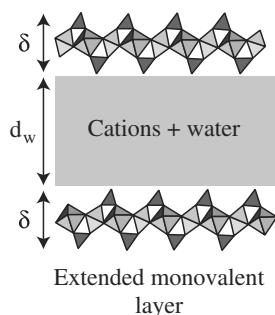


**Fig. 10.28.** Schematic view of a bilayer vesicle based on pentaphenyl fullerene. The outer layer is made up of 6 693 molecules and the inner layer of 5 973 molecules. The outer radius of the vesicle is 17.6 nm. Part of the vesicle has been removed to make the picture easier to interpret

Vesicles are much bigger objects, even ‘giant’, measuring up to 200  $\mu\text{m}$  in diameter for certain assembly processes. Naturally, the phospholipids are not the only compounds which lend themselves to this kind of shaping procedure. For example, vesicles have recently been fabricated with the sodium salt of pentaphenyl fullerene (see Fig. 10.28) [29]. The formation of micelles or vesicles may thus allow a specific molecule, in this case the fullerene, to be placed in solution. Such supramolecular objects can then serve as biocompatible vectors for administering medicines (liposomes) or for carrying out genetic engineering, or as microreactors for specific chemical syntheses. Cell membranes incorporate proteins in their structure. In the same way, molecules carrying various functional groups can be incorporated into vesicles, e.g., groups establishing lock-and-key recognition mechanisms. Interventric interactions are then induced which can be used to create higher-level assemblies.

Mesophases constitute another example of elementary supramolecular self-organisation, well known to us through its widespread use as the active element in liquid crystal displays. These mesophases can be used as hosts for other molecules to which they then transfer their organisation. In certain types of display, these may be dichroic dye molecules, for example, but it has also been possible to orient endohedral fullerenes  $N@C_{60}$  and  $N@C_{70}$  [30]. This





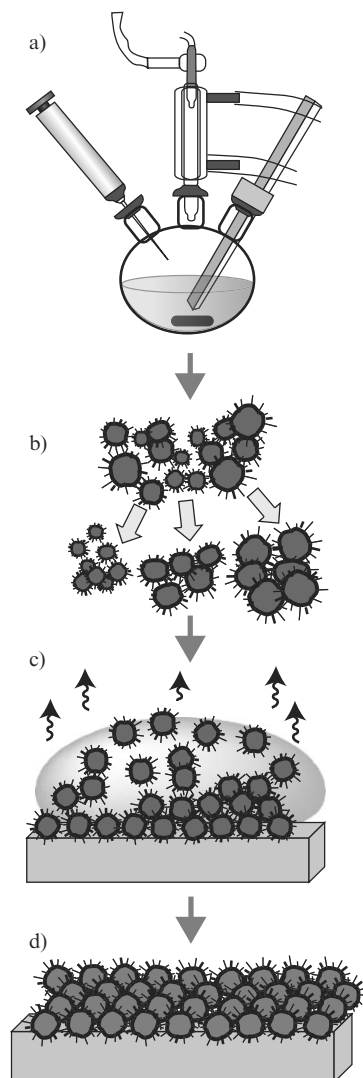
**Fig. 10.29.** Cross-section showing the structure of  $\text{H}_3\text{Sb}_3\text{P}_2\text{O}_{14}$ . Crystal layers of thickness  $\delta$  are regularly spaced and separated by water layers of thickness  $d_w$ . The *dark tetrahedra* are the  $\text{PO}_4$  groups and the *shaded octahedra* are the  $\text{SbO}_6$  groups

interesting observation suggests that one may be able to carry out self-organisation of molecular magnets.

In another field, a recent study [31] has shown that it is possible to grow crystal monolayers of phosphoantimonate in a lyotropic mesophase (see Fig. 10.29). This has several interesting consequences. First, in the crystal layers the atoms are covalently bound and the interlayer spacing can be adjusted over the considerable range 1.5–225 nm. Furthermore, the material can be mechanically or magnetically oriented.

Amphiphilic molecules self-organise under the effects of attractive and repulsive forces due to the hydrophilic and hydrophobic parts. Other types of molecule known as diblock copolymers react to the similar effect of microsegregation to form ordered structures. These macromolecules are formed from the association of two immiscible polymer chains, such as polystyrene (PS) and poly(methylmethacrylate) (PMMA). When such a polymer is cooled slowly from its melt state, a micro-phase separation occurs, leading to the formation of domains containing only PMMA or only PS. Depending on the type of blocks, their size, and the experimental conditions, a wide variety of ordered solid phases can be obtained. This self-organisation ability persists when thin films are fabricated with these materials. They can thus be used as templates for organising other molecules or atoms (see Sect. 10.3.2).

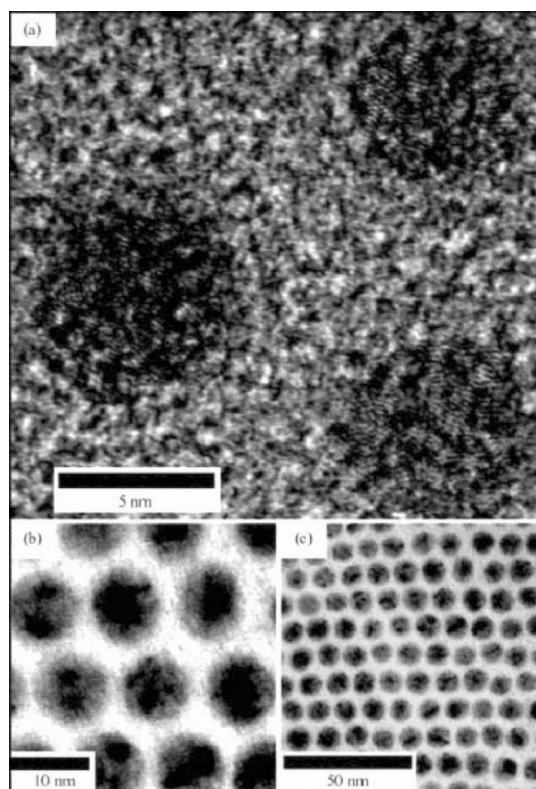
When solid particles are dispersed in a liquid, the result is an unstable suspension which rapidly sediments out. However, if the particles are coated with a surfactant layer, a stable colloidal suspension is obtained. This property is used in a whole range of fields from pharmaceuticals to paint manufacture, not to mention the food industry. Another example is provided by the ferrofluids: when coated with surfactant, magnetic microparticles lose their tendency to cluster under the effects of magnetic forces and form a stable (magnetic) liquid. This can be used in rotary seals, for example. Using this same property, IBM research scientists have been able to fabricate colloidal suspensions of cobalt or iron–platinum nanoparticles (a few nanometers in diameter) which self-organise on surfaces or in the bulk simply by evaporating the solvent [32]. The final result is a 3D superlattice of nanomagnets (see Figs. 10.30 and 10.31). The ultimate aim here is of course to be able to use each of these nanoparticles as a storage element in a magnetic RAM.



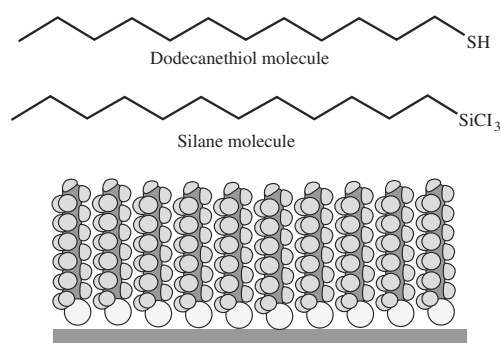
**Fig. 10.30.** Self-organisation of metallic nanoparticles. (a) Synthesis in solution. (b) Choosing the particle size by selective precipitation. (c) Deposition from a suspension. (d) Spontaneous formation of a 3D architecture by evaporating the solvent. A final high temperature annealing stage pyrolyses the surfactant molecules and coats the metal particles with a protective carbon layer. Courtesy of IBM [32]

### 10.3.2 Self-Assembly on Surfaces

All molecules have a natural tendency to prefer the company of identical molecules. This is in fact a first step towards self-assembly, which can be exploited rather easily to make organised monolayers on surfaces. The basic idea is used with molecules like alkanethiols, which spontaneously form dense organised layers on gold or silver surfaces, or silanes, which react with glass or silica surfaces. For example, an alkanethiol solution is prepared in an organic solvent and the metal surface is simply dipped into it. The thiol group fixes on the surface covalently and the alkyl chains form a dense brush (see Fig. 10.32).

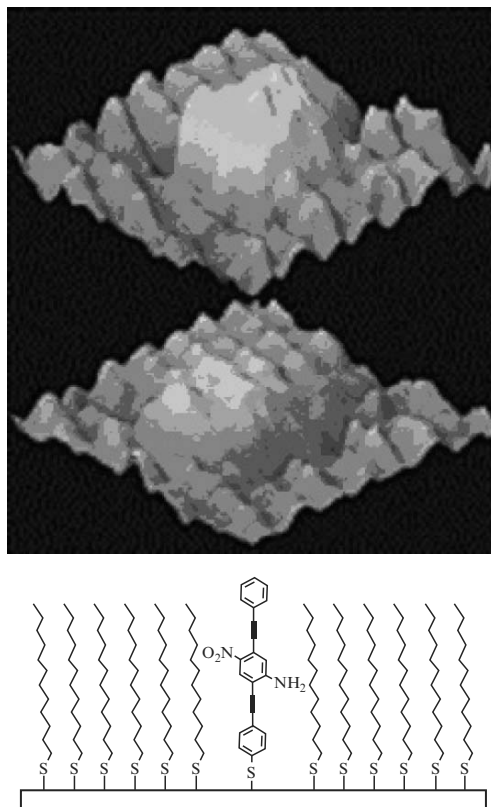


**Fig. 10.31.** TEM images of cobalt nanoparticles with different magnifications. Courtesy of IBM [32]



**Fig. 10.32.** Schematic representation of an alkanethiol monolayer on a gold surface

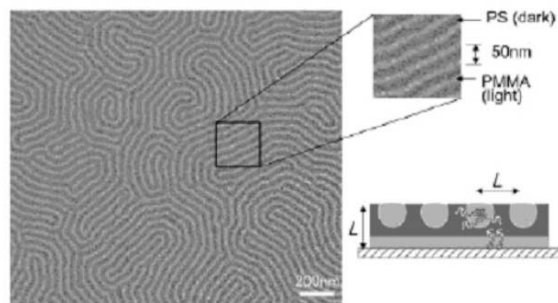
This type of layer has recently been used to demonstrate that the phenylene-ethynylene oligomer can be used as an electrical switch (see Fig. 10.33) [33]. The oligomer, surrounded by insulating dodecanethiol molecules, is itself



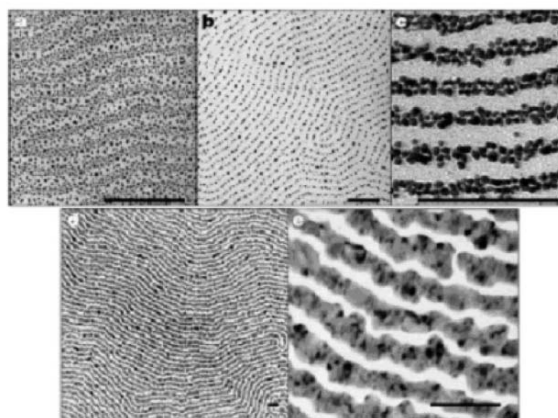
**Fig. 10.33.** *Lower:* Phenylene-ethynylene oligomer isolated in a layer of dodecanethiol. One can act on the conductance of the molecule in such a way that it then retains its conducting or insulating state for several hours, depending on the experimental parameters, and especially the density of the dodecanethiol layer. *Upper:* STM images of an isolated molecule in its conducting state (*top*) and insulating state (*bottom*). The increase in the tunneling current in the first case makes the molecule appear higher. Courtesy of Penn State University [33]

tethered onto the surface by a thiol group. A topographic image of the surface is then obtained using a scanning tunneling microscope. When the molecule is in its conducting state, it appears as a bright point on the STM images. Applying electrical pulses to the STM tip, the experimenters were able to switch the molecule. The stability of the conducting or insulating state depends on the density of the dodecanethiol layer, which seems to suggest that switching is related to the molecular conformation.

As mentioned above, the diblock copolymers spontaneously form organised structures by micro-phase separation. Nanoscale structures can be ‘drawn’ by depositing such materials on a surface (see Fig.10.34) [34]. The authors of this work had the idea of using these structures as a kind of scaffolding for



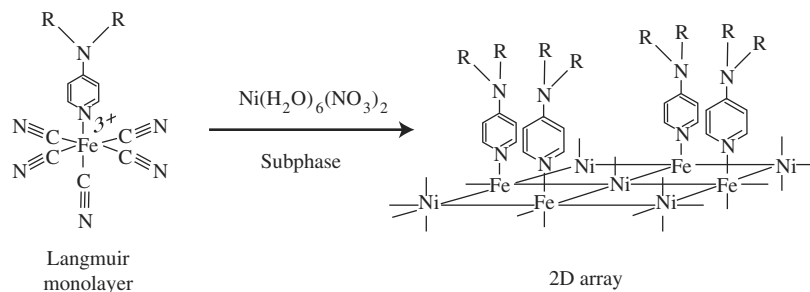
**Fig. 10.34.** TEM image of the nanoarray formed by deposition of a PS–PMMA diblock copolymer on a surface



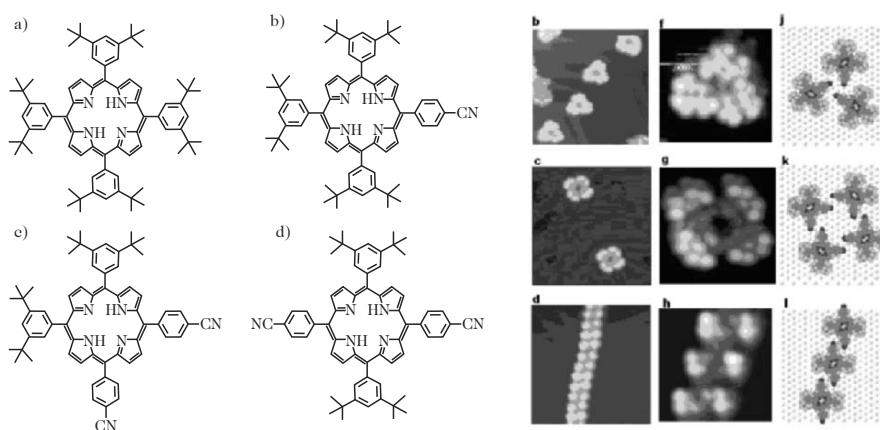
**Fig. 10.35.** Metal nanochains and nanowires formed on PS–PMMA structures. (a) Gold after deposition. (b) Gold after annealing. (c) After iterating the deposition/annealing process. Note the formation of wires by selective migration. (d) and (e) Wide-angle and detailed view of silver wires

building other structures. By playing on the selective wetting properties of these two sorts of polymers (PS and PMMA), they were able to force metal atoms to reproduce the patterns drawn on the surface. Hence, gold or silver atoms prefer the polystyrene lines, whereas lead, indium or tin prefer the PMMA domains (see Fig. 10.35). In this way, one can obtain nanoscale arrays of conducting metal wires. Such wires could be used to connect arrays of functional nanostructures such as transistors, memory cells, or electroluminescent diodes.

The Langmuir–Blodgett technique is a well-tried method for shaping molecular compounds, but always spectacular. Once again the underlying phenomenon is amphiphilicity, although molecules with no amphiphilic tendency have



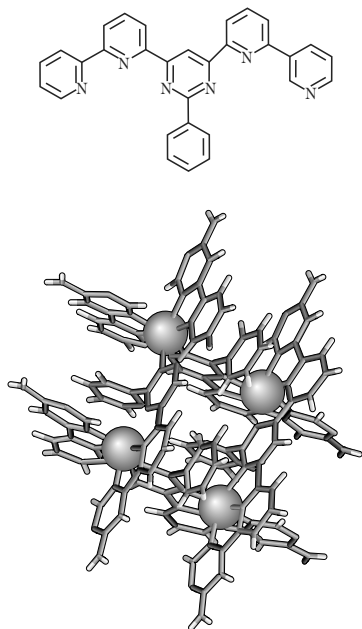
**Fig. 10.36.** Formation of a flat 2D square array using coordination bonds



**Fig. 10.37.** Structures formed by self-assembly on a gold surface by molecules (b), (c) and (d). The molecules (d) spontaneously form strings

also been shaped using this process. Amphiphilic molecules are dissolved in an organic solvent, then deposited on a clean water surface. When the solvent is evaporated, the molecules self-organise with their polar head in the water and their hydrophobic tail in the air. It is easy to manipulate this so-called Langmuir film using a scraper to transfer it to a solid supporting surface. This therefore provides a simple means of fabricating ordered molecular assemblies on surfaces.

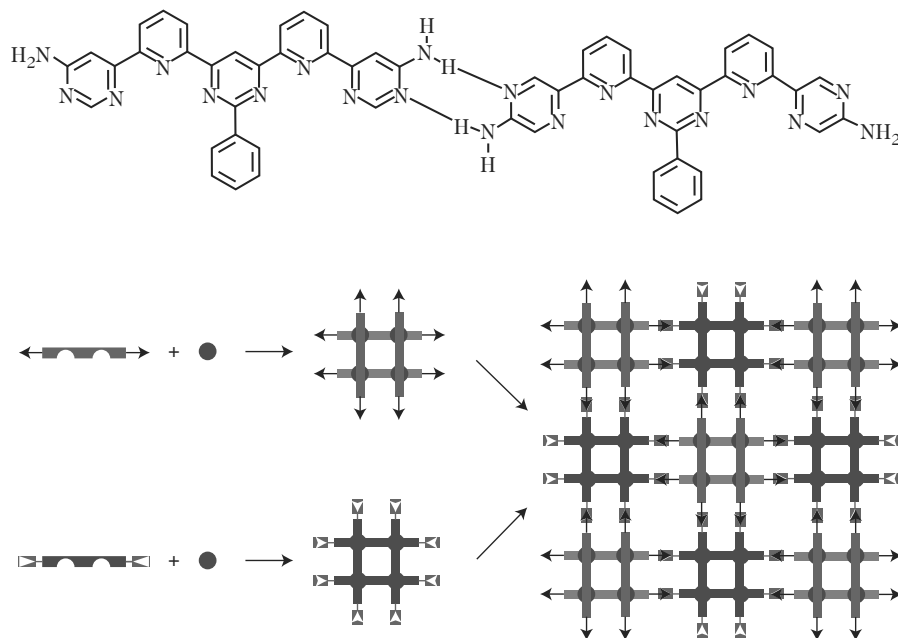
Using this technique, it has been possible to form a flat 2D square array with the molecule shown in Fig. 10.36 associated with nickel nitrate  $\text{Ni}(\text{NO}_3)_2$ . In the resulting films, the iron and nickel atoms are connected by a ligand. These films are easily transferred onto solid supporting surfaces. The important point is that the molecular organisation thereby obtained cannot be ‘naturally’ achieved by conventional chemistry. The square array made here resembles the cyanoferrate Prussian blue. It has similar magnetic properties, but lower dimensionality.



**Fig. 10.38.** Example of a supramolecular grid obtained by self-assembly of four ligands around four metal cations

Generally speaking, in self-assembly processes occurring on surfaces, only substrate–molecule interaction forces are used to form structures. This leads to an array whose density depends on purely steric interactions between neighbouring molecules. Naturally, it is tempting to try to control growth in the plane. By careful molecular design, it is possible to organise the induced structures via anisotropic steric interactions. However, it seems more interesting to exploit forces such as dipolar interactions to impose preferred growth axes. Using this idea, it has been shown that molecules with a strong dipole moment spontaneously form structures whose symmetry depends on the position of the dipole moment (see Fig. 10.37). Moreover, the molecules are preferentially deposited at certain points on the gold  $\langle 111 \rangle$  surface, whereupon one may attempt the directed formation of structures such as conducting wires connecting specific points on the surface.

J.M. Lehn and coworkers regularly produce magnificent examples of self-assembled molecules [4]. Using ligands related to the terpyridine family, it has been possible to fabricate supramolecular grids with metal ions (e.g., iron, cobalt, etc.) at the intersections (see Fig. 10.38). Spin transition complexes can be made if iron atoms are used. These compounds possess two stable spin states, called high spin and low spin, and transitions can be induced between the two states by adjusting an external parameter. In the present case, the material has three spin states and can switch from one to another under the effects of light, pressure or temperature. By suitably functionalising the end of the ligands, the formation of hydrogen bonds can be favoured and



**Fig. 10.39.** Grid of grids

one may then achieve a second level of self-assembly, viz., a grid of grids (see Fig. 10.39). Complexes can be made which incorporate two different metals, thereby producing a chessboard structure. Such assemblies may then exhibit remarkable magnetic or electrochemical properties.

## References

1. Newkome, G.R., Moorefield, C.N., Vögtle, F.: *Dendrimers and Dendrons: Concept, Synthesis, Applications*, Wiley-VCH, Weinheim (Germany) (2001)
2. Fréchet, J.M.J., Tomalia, D.A.: *Dendrimers and Other Dendritic Polymers*, Wiley Series in Polymer Science, Sussex (UK) (2001)
3. Lehn, J.M.: *Supramolecular Chemistry, Concepts and Perspectives*, Wiley-VCH, Weinheim (Germany) (1995)
4. Lehn, J.M.: *Chem. Eur. J.* **6**, 2097–2102 (2000)
5. Dietrich-Buchecker, C.O., Sauvage, J.P., Kintzinger, J.P.: *Tetrahedron Lett.* **24**, 5095–5098 (1983)
6. Solladié, N., Chambron, J.-C., Dietrich-Buchecker, C.O., Sauvage, J.-P.: *Angew. Chem. Int. Ed. Engl.* **35**, 906–909 (1996)
7. Dietrich-Buchecker, C.O., Nierengarten, J.F., Sauvage, J.P., Armaroli, N., Balzani, V., De Cola, L.: *J. Am. Chem. Soc.* **115**, 11237–11244 (1993)
8. Takeda, N., Umemoto, K., Yamaguchi, K., Fujita, M.: *Nature* **398**, 794–796 (1999)



9. Hunter, C.A., Hyde, R.K.: *Angew. Chem. Int. Ed. Engl.* **35**, 1936–1939 (1996)
10. Marvaud, V., Vidal-Ferran, A., Webb, S.J., Sanders, J.K.M.: *J. Chem. Soc., Dalton Trans.* 985–990 (1997)
11. Brettar, J., Gisselbrecht, J.P., Gross, M., Solladié, N.: *Chem. Commun.* 733–734 (2000)
12. Conn, M.M., Rebek Jr., J.: *Chem. Rev.* **97**, 1647–1668 (1997)
13. Sessler, J.L., Wang, B., Harriman, A.: *J. Am. Chem. Soc.* **115**, 10418–10419 (1993)
14. Hayashi, T., Ogoshi, H.: *Chem. Soc. Rev.* **26**, 355–364 (1997)
15. Zimmerman, S.C., Zeng, F., Reichert, D.E.C., Kolotuchin, S.V.: *Science* **271**, 1095–1098 (1996)
16. Timmerman, P., Prins, L.J.: *Eur. J. Org. Chem.* 3191–3205 (2001)
17. Ferlay, S., Felix, O., Hosseini, M.W., Planeix, J.M., Kyritsakas, N.: *Chem. Commun.* 702–703 (2002)
18. Clark, T.D., Buehler, L.K., Ghadiri, M.R.: *J. Am. Chem. Soc.* **120**, 651–656 (1998); Bong, D.T., Clark, T.D., Granja, J.R., Ghadiri, M.R.: *Angew. Chem. Int. Ed. Engl.* **40**, 988–1011 (2001)
19. Harada, A., Takahashi, S.: *J. Chem. Soc., Chem. Commun.* 1352–1353 (1988)
20. Graf, E., Hajek, F., Hosseini, M.W., Planeix, J.M., De Cian, A., Kyritsakas, N., Fischer, J.: *Lettre des Sciences Chimiques* 1–9 (1996)
21. Felder, D., Heinrich, B., Guillon, D., Nicoud, J.F., Nierengarten, J.F.: *Chem. Eur. J.* **6**, 3501–3507 (2000)
22. Hawker, C.J., Wooley, K.L., Frechet, J.M.J.: *J. Chem. Soc. Perkin Trans. 1* **21**, 1287–1297 (1993)
23. Ashont, P.R., Diederich, F., Gómez-López, M., Nierengarten, J.-F., Preece, J.A., Raymo, F.M., Stoddart, J.F.: *Angew. Chem. Int. Ed. Engl.* **36**, 1448–1451 (1997)
24. Balzani, V., Credi, A., Raymo, F.M., Stoddart, J.F.: *Angew. Chem. Int. Ed. Engl.* **39**, 3348–3391 (2000)
25. Bedard, T.C., Moore, J.S.: *J. Am. Chem. Soc.* **117**, 10662–10671 (1995)
26. Bissell, R.A., Cordova, E., Kaifer, A.E., Stoddart, J.F.: *Nature* **369**, 133–137 (1994)
27. Jimenez, M.C., Dietrich-Buchecker, C., Sauvage, J.P.: *Angew. Chem. Int. Ed. Engl.* **39**, 3284–3287 (2000)
28. Eigler, D.M., Schweizer, E.K.: *Nature* **344**, 524–526 (1990)
29. Zhou, S., Burger, C., Chu, B., Sawamura, M., Nagahama, N., Toganoh, M., Hackler, U.E., Isobe, H., Nakamura, E.: *Science* **291**, 1944 (2001)
30. Meyer, C., Harneit, W., Lips, K., Weidinger, A.: *Phys. Rev. A* **65**, 61–201 (2002)
31. Gabriel, J.C.P., Camerel, F., Lemaire, B.J., Desvaux, H., Davidson, P., Batail, P.: *Nature* **413**, 504 (2001)
32. Sun, S., Murray, C.B., Weller, D., Folks, L., Moser, A.: *Science* **287**, 1989 (2000)
33. Donhauser, Z.J., Mantooth, B.A., Kelly, K.F., Bumm, L.A., Monnell, J.D., Stapleton, J.J., Price Jr., D.W., Rawlett, A.M., Allara, D.L., Tour, J.M., Weiss, P.S.: *Science* **292**, 2303 (2001)
34. Lopes, W.A., Jaeger, H.M.: *Nature* **414**, 735 (2001)

## **Part III**

---

### **Properties and Applications**

## Ultimate Electronics

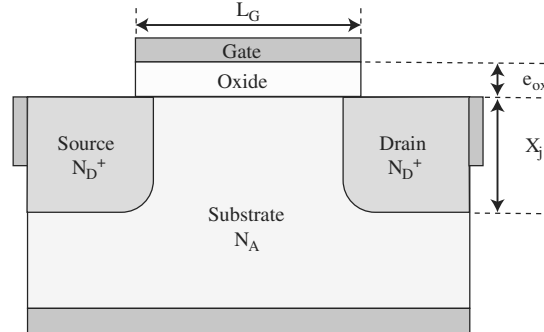
S. Galdin-Retailleau, A. Bournel, and P. Dollfus

The trend in microelectronics is a reduction in the characteristic dimensions of components, with the aim of improving both the integration density and the speed of circuits. The gate length  $L_G$  of MOSFET devices (metal oxide semiconductor field effect transistors) like the one illustrated in Fig. 11.1 has thus been whittled down by progress in lithography from 350 nm ten years ago to about 80 nm today. It is predicted to reach 50 nm in one or two years, and 25 nm in seven or eight years. In integrated circuits based on MOSFETs, the number of transistors per circuit can be greatly increased, in fact, by a factor of about 1.4 each year, close to the famous prediction of Gordon Moore in 1965 [1]. They can thus carry out more and more complex tasks, whilst increasing the operating frequency. However, this trend towards what one might call the ultimate electronics, or nanoelectronics since we shall see that all the characteristic MOSFET dimensions tend towards ten nanometers or less, is going to raise new problems, both technological and physical.

In Sect. 11.1, we outline the main operating principles of CMOS (complementary metal oxide semiconductor) integrated circuits, the aim being to identify the principal design parameters. We shall then review the main scaling rules used up to now to reduce  $L_G$ , without impairing the effective operation of the MOSFET (Sect. 11.2). Indeed, as  $L_G$  is reduced, certain undesirable side-effects show up, which one attempts to counter by adjusting the other geometrical parameters defining the MOSFET structure, in particular, the thickness  $e_{ox}$  of the gate oxide layer, the doping profile of the active region, and the depth  $X_j$  of the source and drain regions. Section 11.3 then deals with alternative MOSFET architectures likely to provide viable industrial solutions to problems that are difficult or even impossible to solve with the conventional architecture shown schematically in Fig. 11.1 for values of  $L_G$  below 50 nm.

### Normally-off MOSFET

The aim of the MOSFET is to control the passage of charge carriers from a source electrode to a drain electrode. To this end, one must:



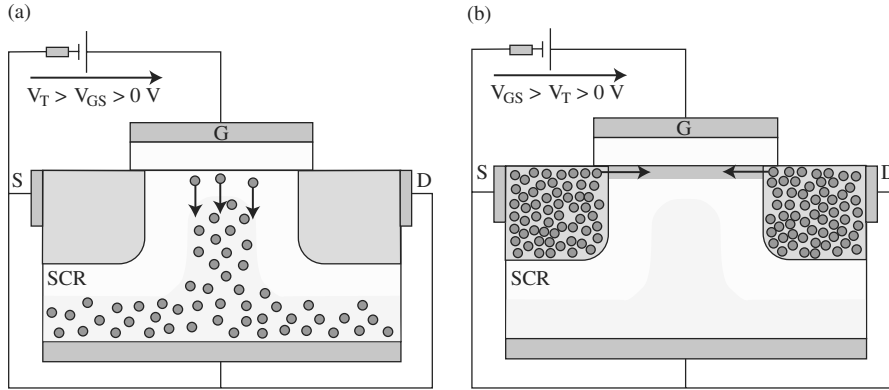
**Fig. 11.1.** Schematic representation of a normally-off electron channel ( $n$ -channel) MOSFET transistor, i.e., in the off state when no bias is applied between the gate and the source. An oxide layer and metallic gate are stacked on a  $p$ -doped substrate with concentration  $N_A$  of acceptors (boron atoms). The thickness of the  $\text{SiO}_2$  layer between the gate and substrate is denoted by  $e_{\text{ox}}$  and the length of the gate by  $L_G$ . Two electron reservoirs, the source and drain regions, highly doped with donors (phosphorous or arsenic, with concentration  $N_D^+$ ), are embedded on either side of the gate. Two electrodes are used to bias the source and drain regions, which are characterised by their depth  $X_j$  in the substrate. A substrate electrode also introduces a bias in the device. The operating principles of this structure are described in the text

1. form a conduction channel between source and drain,
2. set the charge carriers in motion between these two electrodes.

For the  $n$ -channel transistor architecture shown in Fig. 11.1, condition (1) can only be fulfilled by imposing a gate voltage that is strictly positive with respect to the other electrodes. This is why it is referred to as a normally-off transistor. In fact, there is no permanently existing channel. It has to be created electrically via the MOS capacitance. However, when the drain-source bias  $V_{\text{DS}}$  and the substrate-source bias  $V_{\text{BS}}$  are both zero, and when a bias  $V_{\text{GS}} > 0$  V is first applied, the holes initially present under the gate are repelled towards the bottom of the substrate, thereby creating a space charge region (SCR) at the Si/SiO<sub>2</sub> interface, as shown in Fig. 11.2a. It is only for a certain minimum bias  $V_{\text{GS}} = V_{\text{T}}$ , the threshold voltage of the transistor, that an inversion layer appears under the Si gate. This is a layer several nanometers thick, filled with electrons provided by the source and drain regions (see Fig. 11.2b). The theoretical value of  $V_{\text{T}}$  usually considered is that of  $V_{\text{GS}}$  corresponding to an electron concentration in the inversion channel at least equal to the concentration  $N_A$  of acceptor dopants in the  $p$ -type substrate. It can be adjusted by suitably modifying  $N_A$  and the type of gate material.

For  $V_{\text{GS}} > V_{\text{T}}$ , the MOSFET is in the electrically on state, but it remains to see whether condition (2) is fulfilled before a current  $I_{\text{D}}$  can circulate between drain and source. In other words, a bias  $V_{\text{DS}} > 0$  V must be applied to impose an accelerating electric field  $E_{\parallel}$  in the channel, parallel to the Si/SiO<sub>2</sub> interface.

When the voltage  $V_{\text{DS}}$  remains small, i.e., as long as there is an electron channel connecting the source to the drain and the speed  $v_n$  of the electrons increases linearly with the value of  $E_{\parallel}$ , i.e.,  $v_n = \mu_n E_{\parallel}$ , where  $\mu_n$  is the electron mobility, the induced



**Fig. 11.2.** Creation of a conduction channel in a normally-off  $n$ -channel MOSFET device. (a) Hole repulsion under the gate. (b) Formation of a conduction channel between the source and drain electron reservoirs

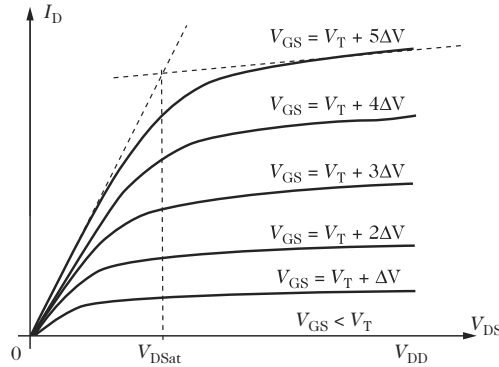
channel behaves like a simple resistor. This is the ohmic regime. For small values of  $V_{DS}$ , the current  $I_D$  can be written to a first approximation in the form

$$I_D = \mu_n C_{ox} \frac{W_n}{L_G} V_{DS} (V_{GS} - V_T),$$

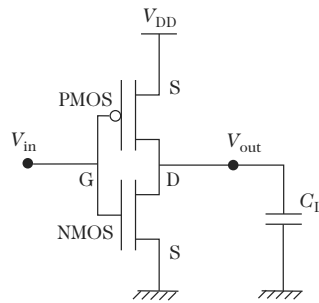
where  $C_{ox} = \varepsilon_0 \varepsilon_{rox} / e_{ox}$  is the surface capacitance of the MOS structure,  $\varepsilon_0$  is the dielectric permittivity of the vacuum,  $\varepsilon_{rox}$  is the relative dielectric permittivity of  $\text{SiO}_2$ , and  $W_n$  is the width of the channel. In the last relation,  $\mu_n V_{DS} / L_G$  corresponds to  $v_n$  and  $C_{ox} (V_{GS} - V_T)$  to the surface charge density in the channel. The ratio  $L_G / [\mu_n C_{ox} (V_{GS} - V_T) W_n]$  represents the resistance of the channel. The mobility  $\mu_n$  is limited by vibrations in the crystal lattice and by collisions between electrons and ionised doping impurities. Note also that, in MOSFET devices, the mobility is significantly hampered in the channel due to the unavoidable roughness of the interface between the monocrystalline material Si and the amorphous material  $\text{SiO}_2$ .

When  $V_{DS}$  goes above the limiting value  $V_{DSsat}$ , known as the pinch-off voltage, the current  $I_D$  no longer increases, or only slightly, with  $V_{DS}$ . This is the saturation region, where the channel depth is reduced to zero and the electron speed saturates [2]. The current  $I_D$  can still be controlled through  $V_{GS}$ . In today's MOSFET devices, the saturation current  $I_{Dsat}$  varies linearly with  $V_{GS}$ . The level of saturation of  $I_D$  is evaluated with respect to  $V_{DS}$  by measuring the drain conductance  $g_D$ , i.e., the slope of  $I_D$  with respect to  $V_{DS}$  at given  $V_{GS}$  for  $V_{DS} > V_{DSsat}$ . Figure 11.3 shows the typical appearance of the characteristic  $I_D(V_{DS})$  for an MOSFET operating according to the principles described above.

In Sect. 11.1, we shall see that, in CMOS technology, a  $p$ -channel normally-off transistor or PMOS is always associated with an  $n$ -channel normally-off transistor or NMOS. To do this, the type of doping is changed in the different regions: the substrate is doped with donors, whereas the source and drain regions are doped with acceptors.  $p$ -channel transistors operate by the same principles as  $n$ -channel transistors, but changing the sign of the biases. In order to satisfy the previously defined conditions (1) and (2), one must apply  $V_{GS} < V_{Tp} < 0$ , where  $V_{Tp}$  is the



**Fig. 11.3.** Typical current–voltage characteristic of a normally-off *n*-channel MOS-FET



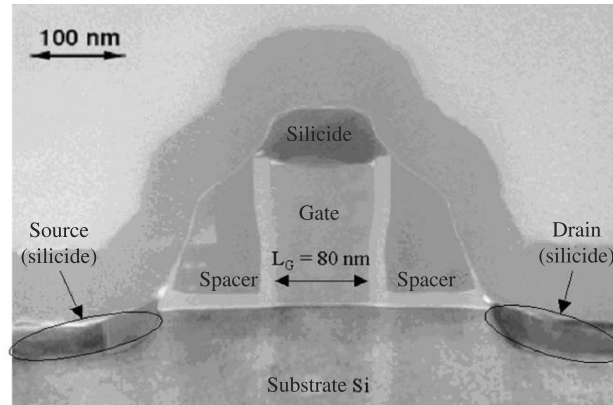
**Fig. 11.4.** Circuit diagram of a CMOS inverter, with  $V_{GSn} = V_{in}$ ,  $V_{GSp} = V_{DD} - V_{in}$ ,  $V_{DSn} = V_{out}$  and  $V_{DSp} = V_{DD} - V_{out}$

PMOS threshold voltage, and  $V_{DS} < 0$ . The NMOS and PMOS can also be made complementary:

- by adjusting  $V_{Tp}$  so that  $V_{Tp} = -V_T$ ,
- by making the PMOS device with width  $W_p$  greater than the width  $W_n$  of the NMOS, so that the currents delivered by these two types of transistor are the same under the same bias conditions. This adjustment is made necessary by the low mobility  $\mu_p$  of holes compared with the mobility  $\mu_n$  of electrons (see Fig. 11.6j).

### 11.1 CMOS Technology

Microelectronics has been dominated for years now by the technology of CMOS digital integrated circuits based on silicon MOSFET devices. This technology associates two types of MOSFET: the electron channel NMOS and the hole channel PMOS, which have complementary operating regimes



**Fig. 11.5.** Transmission electron microscope view of a MOS transistor with gate length 80 nm (taken from [4]). The various types of doping cannot be distinguished on such an image

with regard to the control bias levels. Logical operations can be carried out on two discrete states, where the 0 logic level corresponds to a voltage close to 0 V and the 1 logic level corresponds to a voltage close to the supply voltage  $V_{DD}$  of the circuit (see p. 390 for a more precise definition).

In the case of the basic unit or inverter of the CMOS logic, a PMOS and an NMOS are connected in series between  $V_{DD}$  and the earth, as shown in Fig. 11.4. (The process for making these inverters is summarised below.) The two transistors are controlled by the same gate voltage  $V_{in}$ . When  $V_{in}$  is equal to  $V_{DD}$ , the NMOS is on and the PMOS off. The capacitance  $C_L$  associated with the output node of the inverter (input capacitances of the following logic stages, capacitances associated with the metallic interconnects connecting the inverters to each other or to the outside of the circuit) discharge through the NMOS and the voltage  $V_{out}$  goes to zero. On the other hand, when the voltage  $V_{in}$  is equal to 0 V, the NMOS is off and the PMOS on, whereupon the capacitor  $C_L$  charges up through the PMOS. The output voltage  $V_{out}$  then becomes equal to  $V_{DD}$ . Therefore, this does indeed swap the high and low logic levels between  $V_{in}$  and  $V_{out}$ .

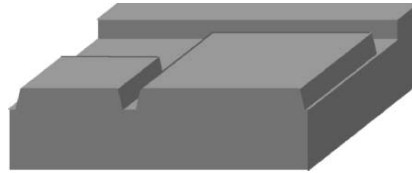
### Fabrication of CMOS Inverters

Figure 11.6 is a schematic representation of the technological processes used to fabricate a CMOS inverter. Details of the various methods can be found in a pedagogical introduction to microelectronics at the website [3].

In order to make this description more concrete, Fig. 11.5 is an electron microscope image of a MOSFET made by STMicroelectronics. It shows the oxide or nitride spacers which serve as a mask at the silicidation stage (see Fig. 11.6j) and procure self-alignment of the silicide contacts established for the source and drain with the silicide contact of the gate.



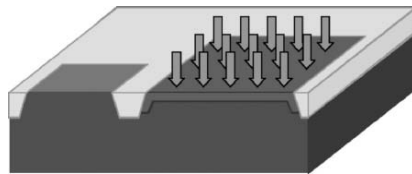
(a) Basic material: *p*-doped Si substrate



(b) Insulating trenches cut by reactive ion etching around the NMOS/PMOS blocks (after defining a resist mask by photolithography)



(c) Vapour phase chemical deposition of an oxide in the trenches, then leveling by mechanochemical polishing



(d) Implantation of donor ions ( $P^-$ , etc.) through a mask to create an *n*-type pseudo-substrate in which the PMOS will be defined. (Implantation also of donors or acceptors to dope the channel: retrograde doping)



(e) Post-implantation anneal to remove defects induced by ion bombardment in the crystal structure and electrically activate the dopants by placing them at substitutional sites. During the anneal, impurities diffuse throughout the substrate



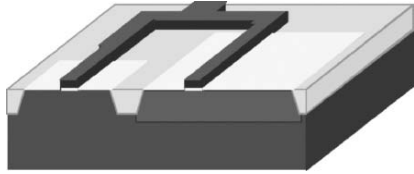
(f) Thermal oxidation of Si to obtain the gate insulator  $SiO_2$

**Fig. 11.6.** Steps in the fabrication of a CMOS inverter. Continued overleaf

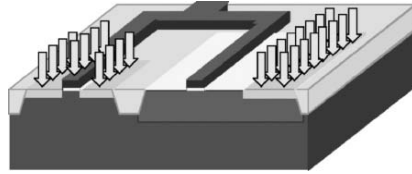
Note that, for a given input logic level, one or two transistors are always in the off state. Apart from the leakage current of these transistors, due to diffusion of the few carriers present under the gate when the channel is not fully formed, no current can circulate in the unit once the voltage  $V_{out}$  has reached one of the two equilibrium output logic levels. To a first approximation, the CMOS inverter does not therefore consume any static power, and this is indeed its main advantage over other technologies for digital integrated circuits.

This brief description of the working of the CMOS inverter is enough to deduce the main parameters characterising the performance of a MOSFET

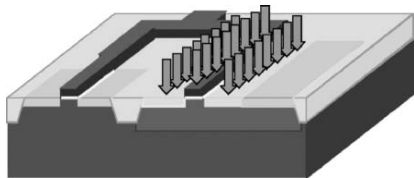




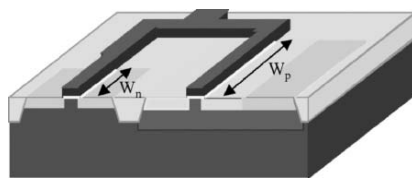
(g) Deposition and reactive ion etching of polysilicon to define gates and their connections



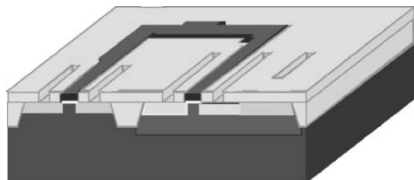
(h) Implantation of donor ions ( $As^-$ , etc.) to create the highly  $n$ -doped regions (source and drain of the NMOS, bias of the  $n$ -type pseudo-substrate). The gate is used as a mask for implanting the source and drain regions, which are then automatically aligned with respect to the gate. One may subsequently use angled ion implantation to fabricate pockets and halos (see Sect. 11.2.3)



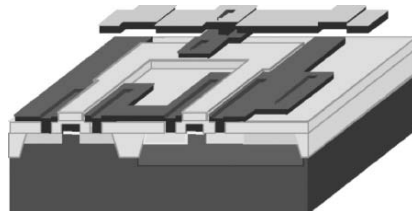
(i) Implantation of acceptor ions ( $B^+$ , etc.) to create the highly  $p$ -doped regions (source and drain of the PMOS, bias of the  $p$ -type substrate)



(j) Post-implantation anneal. The gate material must be able to sustain this heat treatment. Polysilicon is chosen for this reason, being refractory and compatible with silicon processing



(k) Deposition of passivation oxide, etching of the oxide layer and silicidation of the Si in the openings in order to fabricate the source and drain contacts, and also of the gate polysilicon



(l) Passivation and fabrication of the various metal interconnection levels. The passivation oxide between the first and second interconnection levels is not shown

**Fig. 11.6. (cont).** Steps in the fabrication of a CMOS inverter

device for the CMOS technology. In CMOS logic, information is transmitted by the charging or discharging of the capacitances  $C_L$  through one or several transistors, i.e., those in the on state in a logic cell. Of course, the aim is to minimise the charging or discharging time  $\tau_D$  of the capacitors in order to increase the operating frequency of the circuit. In the case of the CMOS inverter, the time  $\tau_D$  typically varies as  $C_L V_{DD}/I_{on}$ , where  $I_{on}$  is the current delivered by the transistor in the on state at the beginning of the change in the charge state of  $C_L$ , i.e., for the voltages  $V_{GS}$  and  $V_{DS}$  equal to  $V_{DD}$  (resp.  $-V_{DD}$ ) for the NMOS (resp. PMOS). The current  $I_{on}$  is thus a determining factor for the speed of the circuit. It must have a value as large as possible. From this point of view, there is every reason to reduce the threshold voltage  $V_T$  in order to increase  $I_D$  at given bias values (see p. 383).

For battery-driven systems, a very low static consumption is clearly essential. For ‘fixed’ systems, it is also essential to maintain circuit heating within reasonable limits. The leakage current  $I_{off}$  crossing an off transistor in equilibrium with the inverter, so that  $V_{GS} = 0$  V and  $V_{DS} = V_{DD}$  (resp.  $-V_{DD}$ ) for the NMOS (resp. PMOS), must therefore be minimised, which a priori means increasing  $V_T$ . This problem is exacerbated as the number of transistors per circuit increases.

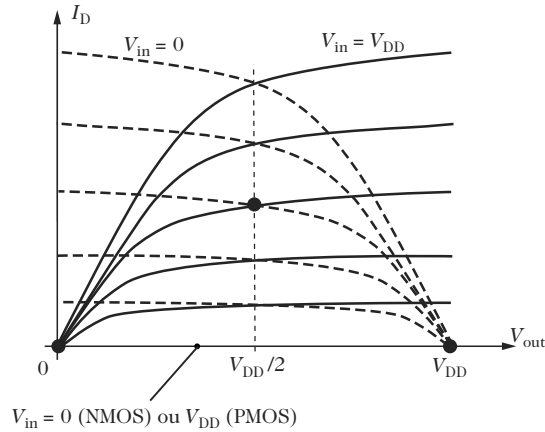
Another issue is to avoid a situation in which a perturbation to the input voltage  $V_{in}$  can alter the output state of the inverter. In order to guarantee that the tolerable noise level does not put too stringent a limit on possible perturbations (see below and esp. Fig. 11.8), the (magnitude of) the threshold voltage  $V_T$  of the transistors must be large enough and the drain conductance  $g_D$  in the saturation region must be as small as possible.

The choice of  $V_T$  thus has to obey conflicting criteria, depending whether one favours dynamic or static performance. A good compromise generally consists in adjusting the value of  $V_T$  in such a way that it is roughly one third the value of  $V_{DD}$ .

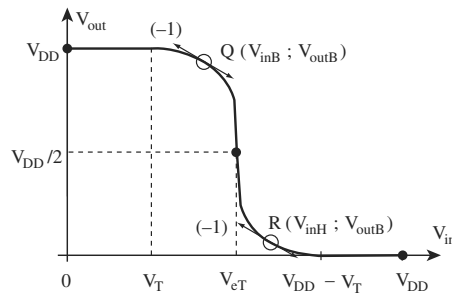
### Input/Output Static Characteristic of a CMOS Inverter

The transfer characteristic  $V_{out}(V_{in})$  of the CMOS inverter is obtained by varying the voltage  $V_{in}$  slowly enough to obtain a quasi-static equilibrium at the output  $V_{out}$ . This therefore corresponds to electrical states in which the currents  $I_{Dn}$  and  $I_{Dp}$  delivered by the NMOS and the PMOS, respectively, are equal.  $V_{out}(V_{in})$  may thus be deduced from the intersections of the characteristics  $I_D(V_{DS})$  of the two transistors in the  $(V_{out}, I_D)$  plane (see Figs. 11.7 and 11.8).

The characteristic  $V_{out}(V_{in})$  shown in Fig. 11.8 reveals 4 zones. For  $V_{in}$  less than  $V_T$ , the NMOS is off, and the PMOS is on and in the ohmic regime (see Fig. 11.7). The output capacitance cannot discharge and  $V_{out}$  remains fixed at  $V_{DD}$ . Then, when  $V_{in}$  varies between  $V_T$  and a limiting value  $V_{eT}$ , the NMOS transistor is unblocked. It operates in the saturation regime, whereas the PMOS tends to move gradually from the ohmic regime to the saturation regime. The voltage  $V_{out}$  decreases as  $V_{in}$  increases. For  $V_{in} = V_{eT}$ , the characteristics of the two transistors cross in the saturation regime. When  $V_{in}$  lies between  $V_{eT}$  and  $V_{DD} - V_T$ , the NMOS goes into



**Fig. 11.7.** Characteristics  $I_D(V_{DS})$  of the NMOS (continuous curves) and the PMOS (dashed curves) in a CMOS inverter, represented in the  $(V_{out}, I_D)$  plane



**Fig. 11.8.** Transfer characteristic  $V_{out}(V_{in})$  of a CMOS inverter

the ohmic regime and the PMOS remains in the saturation regime, whereupon  $V_{out}$  continues to fall. Finally, the PMOS goes off for  $V_{in}$  greater than  $V_{DD} - V_T$  and  $V_{out}$  reaches 0 V.

The input voltage  $V_{eT}$  made equal to  $V_{DD}/2$  by matching the NMOS and PMOS devices, i.e., adjusting the threshold voltages and widths of the transistors, is the switching threshold of the inverter. The variation of  $V_{out}$  near  $V_{eT}$  becomes all the more abrupt as the drain conductance  $g_D$  in the saturation regime is low. For  $g_D = 0$ , a vertical transition occurs in  $V_{out}$  when  $V_{in} = V_{eT}$ , the characteristics  $I_D(V_{DS})$  of the two transistors scanning in saturation regime, with an infinite number of possible values of  $V_{out}$  for this single value of  $V_{in}$ . There is every reason to move towards this ideal limit, which enhances regeneration of the logic levels.

## 11.2 MOSFET Scaling

### 11.2.1 Basic Principles

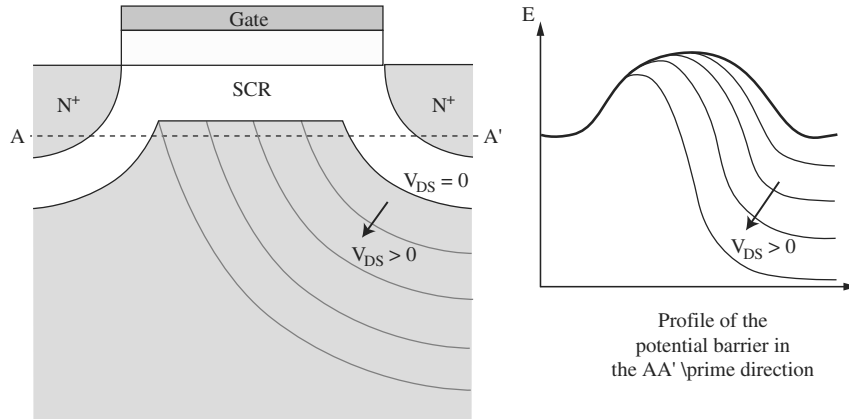
In the last section, it was explained that, in order to improve CMOS circuit speeds, the current  $I_{\text{on}}$  delivered by a MOSFET in its on state needs to be increased. The strategy adopted up to now to achieve this end has always been to gradually reduce the gate length  $L_G$ . Indeed, the resistance due to the channel itself when it is formed then decreases as  $L_G$  decreases, which induces an increase in the drain current for a given bias. Moreover, reducing the size of the device is advantageous as far as packing density is concerned within an integrated circuit.

If  $L_G$  is reduced to increase  $I_{\text{on}}$ , care must be taken to make sure that the current  $I_{\text{off}}$  in the off state and the drain conductance  $g_D$  in the saturation regime are maintained at acceptable levels. The ratio  $V_T/V_{DD}$  must also be carefully controlled. However, these goals can only be achieved if well specified scaling rules are respected for MOSFET devices. Reducing  $L_G$  involves altering the other parameters specifying the transistor geometry ( $\epsilon_{\text{ox}}$ ,  $X_j$ , substrate doping, etc.). As we shall now see, these rules consist for the main part in maintaining capacitive control by the gate of the formation of the conduction channel between source and drain.

### 11.2.2 Short Channel Effects

When  $L_G$  is reduced, electrostatic interference effects begin to appear, known as short channel effects, which perturb the precise control via  $V_{GS}$  of the conductivity between source and drain.

- When the drain is brought near the source, the drain–substrate and source–substrate space charge regions (SCR) are also brought closer together. Normally, these SCRs represent an obstacle to the current moving towards the substrate. The majority carriers of the source and drain regions see a potential barrier which prevents them from diffusing towards the substrate (and conversely for the majority carriers in the substrate which might otherwise move towards the source and drain regions). When the voltage  $V_{DS}$  increases, the drain–substrate SCR extends and can, for small  $L_G$ , join up with the source–substrate SCR. The potential barrier at the source–substrate boundary then falls off, as illustrated in Fig. 11.9. The majority carriers in the source can under these conditions diffuse into the substrate and then drift towards the drain in the oppositely polarised drain–substrate SCR. A leakage current between source and drain, which is a diffusion current not controlled by the gate, thus transits via the substrate. This phenomenon is known as bulk punch-through.
- Furthermore, the advance of the source–substrate and drain–substrate SCRs under the gate increases in relative value compared with  $L_G$  when



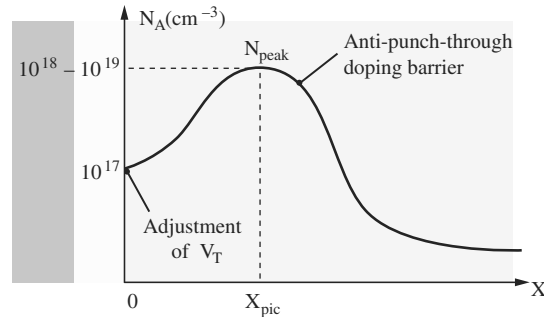
**Fig. 11.9.** Illustration of the bulk punch-through phenomenon in the case of an NMOS in the off state

the value of the gate length decreases. This leads to a lowering of the potential barrier at the channel entrance and perturbs the control via  $V_{GS}$  of fixed charges in the depletion zone under the gate. As a consequence, a surface punch-through phenomenon occurs, analogous to the bulk punch-through effect described previously, the drain conductance  $g_D$  increases in the saturation regime, and finally, the (magnitude of the) threshold voltage  $V_T$  falls (threshold voltage roll-off). These phenomena are exacerbated as (the magnitude of)  $V_{DS}$  increases.

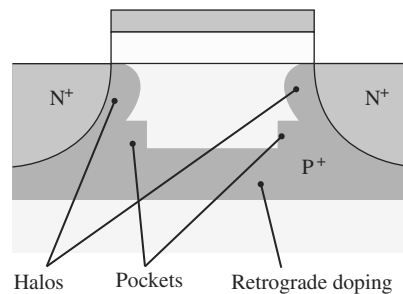
The short channel effects discussed here thus tend to impair control by the gate and significantly increase the current  $I_{off}$  and drain conductance  $g_D$  in the saturation regime, as well as inducing a dependence of  $V_T$  on  $V_{DS}$ . These consequences can be catastrophic for the effective operation of a CMOS circuit. In Sect. 11.2.3, we shall examine the various scaling rules which allow the MOSFET devices to continue to operate correctly when their gate lengths are reduced. In Sect. 11.2.4, we shall see how these rules turn out from a quantitative point of view.

### 11.2.3 Scaling Rules

A relatively simple solution for reducing the consequences of short channel effects is to make a global increase in the substrate doping when  $L_G$  decreases, in order to limit extension of the source–substrate and drain–substrate SCRs. However, this solution soon reaches its limits because it has a significant influence on the value of  $V_T$  and, even more importantly, it has harmful consequences for the mobility of charge carriers in the channel due to the increase in the number of collisions between carriers and ionised impurities. Retrograde doping architectures are preferred, as illustrated in Fig. 11.10. The doping



**Fig. 11.10.** Typical retrograde doping profile for the substrate in the  $x$  direction perpendicular to the metal–oxide–semiconductor stack



**Fig. 11.11.** Architecture of an NMOS with retrograde doping and halos

level at the oxide/Si interface is chosen to obtain the desired value of  $V_T$  and remains at a reasonable level, so as not to seriously impair mobility in the channel, and a buried layer is introduced by implantation or epitaxy over the whole length of the channel with a concentration of dopant (of the same type as in the substrate) that is much higher, designed to limit short channel effects. In order to introduce a further degree of freedom into the transistor design, one may also implant so-called pockets or halos with very high doping levels in the vicinity of the source and drain regions (see Fig. 11.11) [5].

The depth  $X_j$  of the source and drain regions is another important parameter for controlling short channel effects. Indeed,  $X_j$  must be reduced at the same time as  $L_G$  so as to limit the increase in the leakage current due to bulk punch-through. A low value of  $X_j$  also helps to check  $V_T$  roll-off due to the short channel effect. However, apart from the technological difficulties in realising this solution, it may also have a detrimental effect on the on-state electrical performance of the MOSFET. Reducing  $X_j$  increases the resistances  $R_{acc}$  affecting access to the channel through the source and drain regions. For large  $L_G$ , the channel resistance  $R_{chan}$  is much bigger than  $R_{acc}$  and in this case a reduction in  $X_j$  has no effect on  $I_{on}$ . However, when  $L_G$  is reduced,  $R_{chan}$  falls off until it becomes of the same order of magnitude as  $R_{acc}$ . The

potential then drops off in the source and drain regions, reducing the effective voltage applied across the channel, whereupon  $I_{\text{on}}$  falls off significantly. To avoid this scenario, a solution is to increase the doping levels in the source and drain regions when  $L_G$  is reduced. One can also improve the ohmic nature of the metal/Si contacts by introducing silicides of Ti or Co at this interface. Another answer is to raise the source and drain regions with respect to the level of the oxide/channel interface.

Finally, the thickness  $e_{\text{ox}}$  of the gate oxide is perhaps the most important parameter to be controlled in MOSFET scaling theory. Since gate control of the channel conductivity becomes more and more seriously affected by parasitic effects due to the drain as  $L_G$  is reduced, it is useful to improve capacitive control in order to counter short channel effects. The solution used up to now to increase the oxide capacitance  $C_{\text{ox}} = \epsilon_0 \epsilon_{\text{rox}} / e_{\text{ox}}$  has been to reduce the thickness  $e_{\text{ox}}$  of  $\text{SiO}_2$  in proportion to  $L_G$ . The ratio  $L_G / e_{\text{ox}}$  thus varies between 40 and 50 in today's CMOS circuits [6]. However, reducing  $e_{\text{ox}}$  means that the supply voltage  $V_{\text{DD}}$  of the circuits must also be reduced, to avoid the risk of breakdown in the oxide. (The electric field in the oxide is at most  $V_{\text{DD}} / e_{\text{ox}}$ .) This reduction of  $V_{\text{DD}}$  also allows one to control the dynamic power dissipation, which tends to increase with the operating frequency of the circuit and the number of transistors in the circuit, hence with the reduction of  $L_G$ . (For a CMOS inverter, the power dissipated per charge/discharge cycle of the output capacitance, of length  $T$ , varies as  $C_L V_{\text{DD}}^2 / 2T$ .)

#### 11.2.4 State of the Art: ITRS Roadmap

In order to summarise the scaling rules that we have been discussing and to indicate the corresponding orders of magnitude, Fig. 11.12 shows graphs of (a) the  $\text{SiO}_2$  thickness  $e_{\text{ox}}$ , (b) the depth  $X_j$  of the source and drain regions, (c) the depth  $X_{\text{peak}}$  of the concentration peak  $N_{\text{peak}}$  of retrograde doping (see Fig. 11.10), and (d)  $N_{\text{peak}}$  itself, as functions of the technological nodes characterising the development of CMOS circuits (here, the nodes between 250 nm and 22 nm). These values result from predictions made periodically in the well known roadmap as a guideline for the coming 10–15 years in order to continue to design faster circuits obeying Moore's law. This roadmap was originally defined on a purely North American basis by the US Semiconductor Industry Association (SIA). Since 1999, it has been drawn up on a worldwide basis and it is now known as the ITRS roadmap (International Technology Roadmap for Semiconductors) [7]. A quantity called the half pitch is associated with each technological node. This is equal to half the minimal separation between two polysilicon or metal lines on the circuit. The corresponding metallurgical (printed) gate length  $L_G$  is typically 1.3–1.5 times smaller than the half pitch for technological nodes greater than or equal to 100 nm (this node was reached in 2003 with  $L_G$  of the order of 65 nm), and 1.7–1.8 times smaller for nodes less than or equal to 90 nm (node reached in 2004 with  $L_G$  of the order of 50 nm).

As explained in the last section,  $e_{\text{ox}}$  follows a roughly linear trend as a function of the technological node (see Fig. 11.12), but with different slopes for the 1997 SIA roadmap (grey circles) and the 2002 ITRS roadmap (black diamonds). Note also that, for the technological nodes between 130 nm and 65 nm, the values of  $e_{\text{ox}}$  indicated by ITRS 2002 are systematically lower than those predicted by SIA 1997. These changes are symptomatic of the industrial acceleration in the reduction of  $e_{\text{ox}}$ , an aggressive trend caused by the stiff competition between integrated circuit manufacturers [8]. This tendency seems to have reached an end with the 50 nm node, at which point the SIA 1997 and ITRS 2002 predictions roughly coincide. The recommended values of  $e_{\text{ox}}$  are then less than 1 nm, i.e., less than 4 atomic layers of  $\text{SiO}_2$ . We shall see in Sect. 11.3.1 that these ultrathin thicknesses of  $\text{SiO}_2$  may constitute a physical limit beyond which it will be difficult to proceed with the reduction of MOSFET dimensions, unless some technological leap is conceived in the fabrication of MOS capacitance or global transistor architecture.

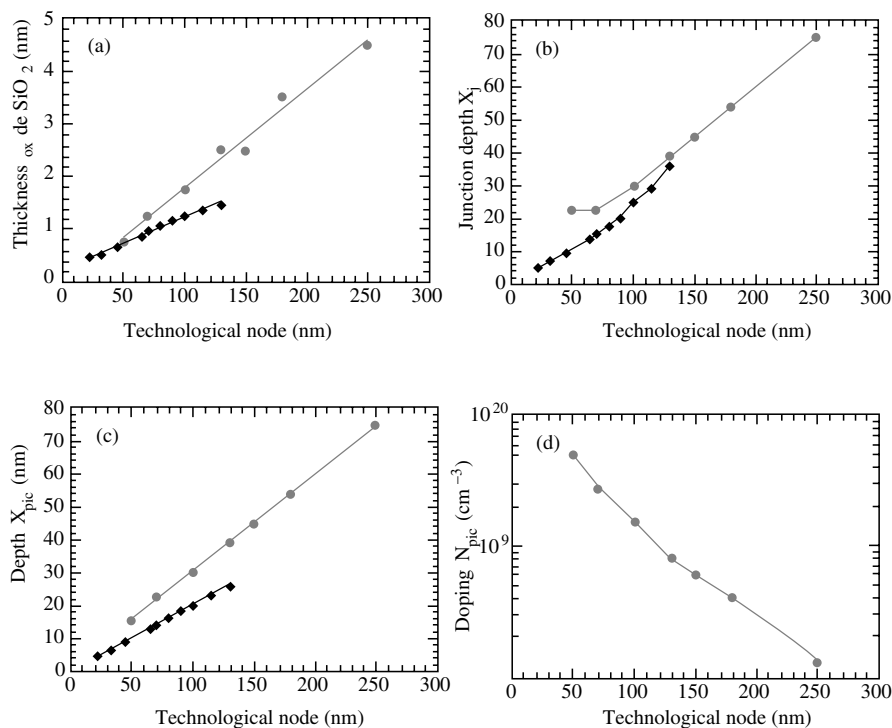
According to the 1997 SIA roadmap, the depth  $X_j$  of the source and drain regions should decrease, like  $e_{\text{ox}}$ , in a roughly linear manner with the technological node, at least down to the 100-nm node, and then for smaller nodes, the value of  $X_j$  tends to saturate at values of the order of 20 nm. At these sizes, given that one seeks to obtain ever more highly doped source and drain regions bounded by highly vertical sidewalls, one does indeed reach the limits of the classical process for doping by implantation followed by annealing described on p. 387. The 2001 ITRS roadmap is for its part more optimistic. Using the new doping methods currently under development, and to which we shall return in Sect. 11.3.1, ever smaller source and drain thicknesses become feasible, with values of  $X_j$  below 10 nm for nodes smaller than 50 nm.

The linear reduction of  $X_j$  with the decrease in the technological node naturally leads to the same effect in the depth  $X_{\text{peak}}$  of the retrograde channel peak, as can be seen from Fig. 11.12c. In the case of the reduction of  $X_{\text{peak}}$  with the technological node, note that, as for  $e_{\text{ox}}$ , the 2002 ITRS roadmap is more aggressive than the 1997 SIA roadmap, and that the reduction of  $X_{\text{peak}}$  would appear to continue without limit.

Figure 11.12d shows the increase in the retrograde doping peak  $N_{\text{peak}}$  when the technological node decreases, an increase that is needed to counter the more and more sensitive short channel effects. This data, provided by the 1997 SIA roadmap, shows that the  $N_{\text{peak}}$  should exceed  $10^{19} \text{ cm}^{-3}$  for nodes below 100 nm. For the same range of technological nodes,  $X_{\text{peak}}$  should be less than 20 nm. This retrograde doping must also be bounded by extremely vertical sidewalls in order to be effective. These remarks are also valid for the fabrication of pockets and halos. Given the difficulties discussed above in carrying out source and drain doping, one may ask whether it will be possible to pursue such methods as a way of countering short channel effects.

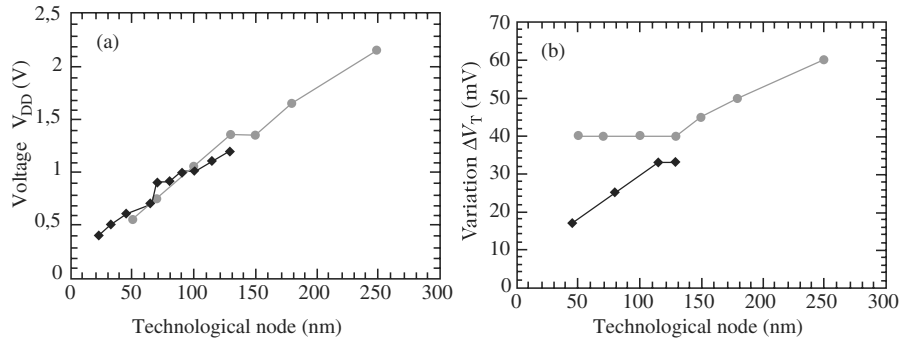
Roadmap predictions are not only concerned with the geometry and doping levels characterising MOSFET architecture, but also with certain electrical parameters. Figure 11.13a shows how the supply voltage  $V_{\text{DD}}$  of CMOS circuits





**Fig. 11.12.** MOSFET scaling trends as a function of technological node according to the 1997 SIA roadmap (*grey circles*) and the 2002 ITRS roadmap (*black diamonds*). (a) Thickness  $e_{ox}$  of the gate insulator  $\text{SiO}_2$ . (b) Depth  $X_j$  of the junction at the interface between the channel and the source and drain regions. (c) Depth  $X_{peak}$  of the retrograde doping peak  $N_{peak}$ . (d) Retrograde doping peak  $N_{peak}$ . Values for  $N_{peak}$  are no longer available in roadmaps after 1997

is expected to evolve as a function of the technological node. Although it is possible to make out steps corresponding to successive standards, the voltage  $V_{DD}$  decreases roughly linearly with the node value. Note also that the 2002 ITRS roadmap is not more aggressive than the 1997 SIA roadmap. Indeed, variations in  $V_{DD}$  are not only related to MOSFET scaling, but also to more or less external factors related to the use of the MOSFET in a CMOS circuit. Furthermore, it can be seen that, for the 100-nm node, the value of the supply voltage  $V_{DD}$  is equal to 1 V, which increases the sensitivity of the technology to fluctuations in the threshold voltage, whose maximal values are indicated in Fig. 11.13b.



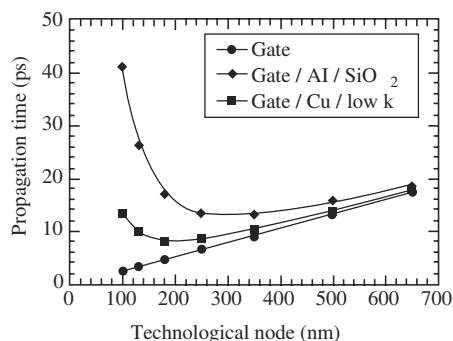
**Fig. 11.13.** MOSFET scaling trends as a function of technological node according to the 1997 SIA roadmap (*grey circles*) and the 2002 ITRS roadmap (*black diamonds*). (a) Supply voltage  $V_{DD}$  and (b) tolerable variations  $\Delta V_T$  in the threshold voltage

### 11.2.5 Interconnects

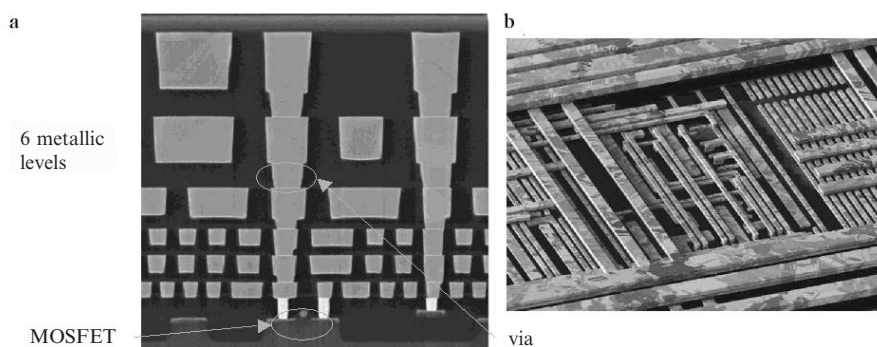
By virtue of the panoply of technological solutions discussed above, it has up to now been possible to continuously reduce the gate length  $L_G$  and increase the current  $I_{on}$  without degrading the overall electrical performance of MOSFET devices. However, it should be said that this improvement in the intrinsic performance of elementary components in the CMOS circuit does not necessarily imply a gain in switching speed such as one might expect from the analysis of the CMOS inverter carried out in Sect. 11.1.

Indeed, the reduction in transistor dimensions requires an appropriate reduction in the cross-section of metallic interconnects joining them together or to their surroundings. These interconnects, separated from one another by insulators, are also closer and closer together as  $L_G$  is reduced. The step from one technological node to a smaller one is thus accompanied by an increase in the resistance  $R_i$  and also the parasitic capacitances  $C_i$  of the interconnects. This in turn implies an increase in the data transmission delay  $R_i C_i$  between logic gates within the integrated circuit, or towards the outside of the circuit. If  $R_i C_i$  grows bigger than the switching time  $\tau_D$  of a gate, the interconnects will become a bottleneck for data transfer.

This problem is illustrated in Fig. 11.14, which compares the dependence of the propagation time  $\tau_D$  on the technological node. The curve plotted through bullets shows the propagation time due to a single logic gate, whilst the curves plotted through diamonds or squares show the propagation time due to a logic gate and its associated interconnects. When the interconnects are made from Al and insulated from one another by  $\text{SiO}_2$  (diamonds), it is observed that, despite the linear decrease in  $\tau_D$  which accompanies the decrease in technological node, the overall propagation time tends to rise significantly with the reduction of the node for nodes smaller than 250 nm. In order to continue to enhance the frequency performance of CMOS circuits, one requires a better



**Fig. 11.14.** Propagation time as a function of technological node. The calculations only account for the delay due to the logic gate (*bullets*), the logic gate and propagation through Al interconnects insulated by SiO<sub>2</sub> (*diamonds*), and the logic gate and Cu interconnects insulated by a low-k insulator (*squares*) (SIA 1997). The interconnects considered here are 0.8  $\mu\text{m}$  thick and 43  $\mu\text{m}$  long



**Fig. 11.15.** Images of Cu interconnects in Intel integrated circuits. (a) Cross-sectional view. (b) Top view. [www.Intel.com/research/silicon/](http://www.Intel.com/research/silicon/)

conductor than Al and a low-k insulator with lower dielectric permittivity than SiO<sub>2</sub>, such as fluorine- or carbon-doped silicon oxide. This is why integrated circuit manufacturers have started to introduce Cu interconnects since 1997–8. These copper interconnects have resistivity  $\rho_{\text{Cu}} = 1.7 \mu\Omega \text{ cm}$  as compared with the previously used aluminium interconnects which had resistivity  $\rho_{\text{Al}} = 3 \mu\Omega \text{ cm}$ . Copper had been avoided until then due to the problems of electromigration posed by this metal. Its use in CMOS processes was made possible by the recent development of barriers of type W<sub>x</sub>N, TaN, etc., which serve to limit diffusion.

In order to be able to cope more easily with the interconnect problem, microelectronics manufacturers also tend to increase the number of metallic levels, which may be as many as 6 or 7 today (see Fig. 11.15). Note, however,

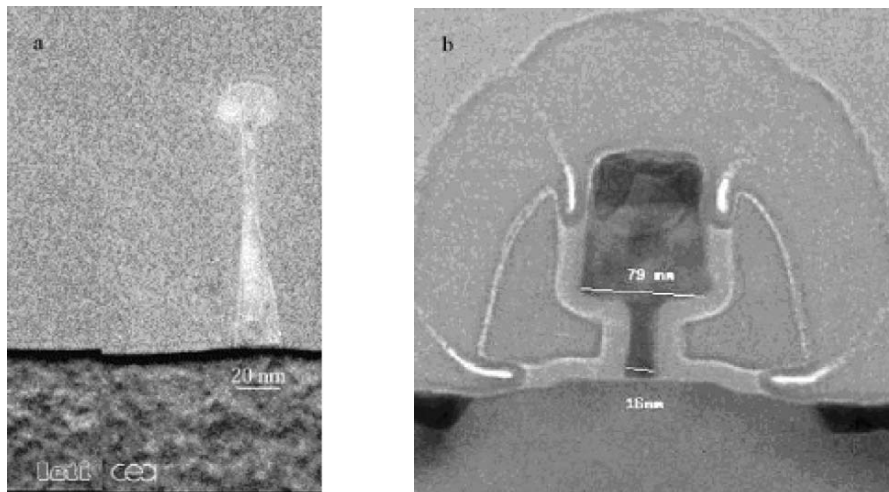
that all the solutions discussed here serve only to postpone the time when interconnects will become a serious obstacle to the move towards faster integrated circuits.

## 11.3 NanoMOS Devices

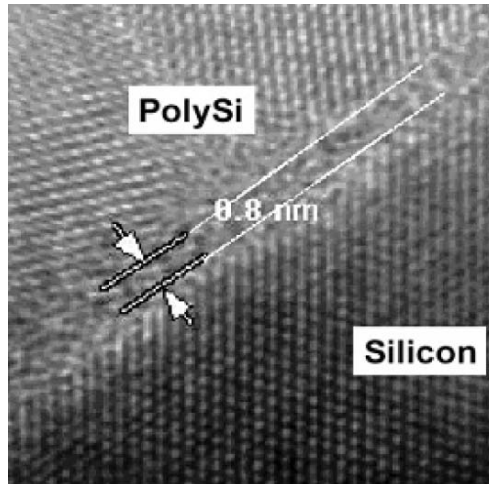
### 11.3.1 Specific Problems

We saw in the last section that, for nodes below 50 nm, all geometric quantities defining the architecture of elementary components in the CMOS technology are brought down to around 10 nm or less. This justifies the term nanoMOS, since a few atoms more or less in one, two or three directions can make a difference. These nanoMOS devices, which already exist in several public and private research establishments (see Fig. 11.16), raise new problems due to their small dimensions, or rehabilitate former difficulties previously overcome or concealed.

Apart from problems of lithography, not considered in this chapter, one of the main technological challenges to be taken up in order to carry this generation of components to the industrial level is to control the growth of a gate insulator that can be used in nanoMOS devices. As discussed above, the ITRS roadmap recommends  $\text{SiO}_2$  thicknesses  $e_{\text{ox}}$  of the order of a few atomic layers, i.e., less than 1 nm. Although MOSFET devices with gate lengths  $L_G$



**Fig. 11.16.** Transmission electron microscope images of different nanoMOS devices. (a) NanoMOS produced by LETI (CEA, Grenoble, France),  $L_G = 20$  nm [9]. (b) NanoMOS produced by STMicroelectronics,  $L_G = 16$  nm [10]



**Fig. 11.17.** High resolution transmission electron microscope view of a polysilicon/SiO<sub>2</sub>/Si stack with SiO<sub>2</sub> thickness equal to 0.8 nm [11]. Dots on the image correspond to electron clouds around atoms

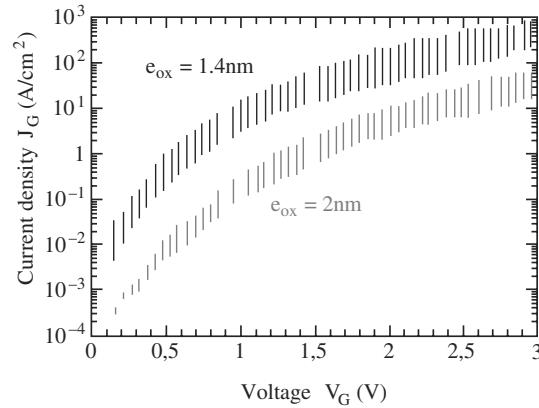
equal to 30 nm and SiO<sub>2</sub> thicknesses equal to 0.8 nm have already been reported (see Fig. 11.17), such thin layers of insulator raise serious problems for the operation of CMOS units.

The main difficulty with such small values of  $e_{ox}$  resides quite simply in controlling the gate insulator layer to within one atomic layer, in as uniform a manner as possible over the whole surface of the channel, not only for a transistor, but also, and perhaps especially, for the tens to hundreds of millions of transistors currently assembled into a typical integrated circuit. There are several reasons why such tight control is essential:

- To obtain reasonably homogeneous electrical characteristics for all transistors in a given chip, and in particular, a low statistical scatter in the threshold voltages  $V_T$ .
- To avoid the appearance of weak points in places where the insulator is thinnest.
- To limit the detrimental effects caused by the roughness of the oxide layer on carrier mobility in the channel.

None of these problems due to technological fluctuations of the SiO<sub>2</sub> are specific to ultrathin layers, but they become increasingly relevant as the mean thickness  $e_{ox}$  is reduced.

The small thickness of SiO<sub>2</sub> also makes it possible for dopants implanted in the gate polysilicon to penetrate into the gate insulator or even into the Si substrate. This phenomenon, due to ion diffusion during post-implantation anneals, is particularly important with boron, a small atom in P<sup>+</sup> doped gates

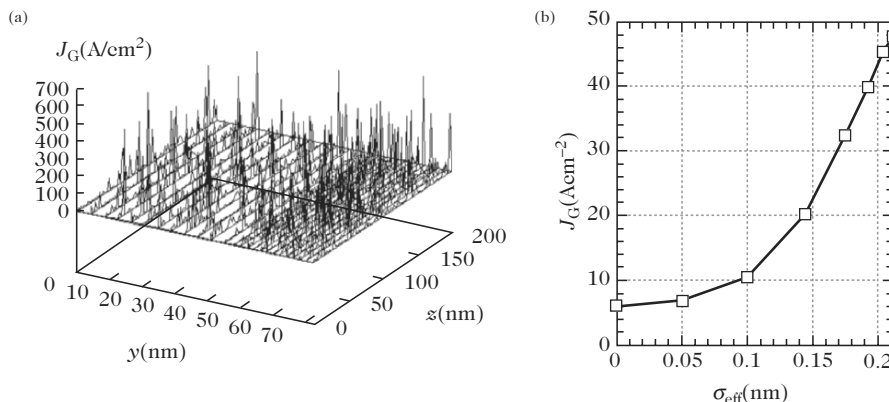


**Fig. 11.18.** Gate current density  $J_G$  as a function of the voltage  $V_G$  applied to this electrode for two MOS capacitances with  $\text{SiO}_2$  thicknesses equal to 2 nm and 1.2 nm [13]. The gate material is highly  $n$ -doped polysilicon (concentration  $5 \times 10^{19} \text{ cm}^{-3}$ ) and the capacitances are made on an  $n$ -type Si substrate with doping concentration  $10^{17} \text{ cm}^{-3}$ . Bars indicate fluctuations obtained over a large number of a priori identical samples

of PMOS devices [12]. It can have extremely harmful effects on the reliability of the oxide layer and on electrical characteristics of transistors such as mobility, threshold voltage, etc.

Finally, a nanometric thickness of  $\text{SiO}_2$  makes it possible for carriers to pass between the gate and channel by the direct tunnel effect. The gate current of the MOSFET is then nonzero, and this constitutes a fundamental modification of the electrical characteristics of this type of device. Figure 11.18 shows measured variations in the gate current density  $J_G$  as a function of the gate voltage for two MOS capacitances with oxide thicknesses  $e_{\text{ox}}$  equal to 2 nm and 1.4 nm [13]. It can be seen that  $J_G$  increases by one or two factors of ten when  $e_{\text{ox}}$  falls by a mere factor of 1.4. This exponential dependence of  $J_G$  on  $e_{\text{ox}}$  is typical of the tunnel current.

The gate tunnel current can have several harmful consequences for the operation of a MOSFET, or more generally, for a CMOS logic cell. To begin with, it can perturb the current  $I_{\text{on}}$ , since channel electrons can escape via the gate before reaching the drain. It can also cause an increase in the current  $I_{\text{off}}$  and hence in the static power dissipated. Finally, it can induce a degradation of the gate oxide by charge injection. During the 1990s, a Japanese group made detailed studies of the operating characteristics of MOSFET devices with gate length  $L_G = 0.1 \mu\text{m}$  and oxide thickness  $e_{\text{ox}} = 1.5 \text{ nm}$  (see [14] and references therein concerning previous work by this group). Their results show that, despite a tunnel current density of a few  $\text{A}/\text{cm}^2$ , such devices perform extremely well in terms of both speed and static power consumption. This can be explained partly by a very good value of the factor  $L_G/e_{\text{ox}}$ , and



**Fig. 11.19.** Effect of non-uniformity of the gate oxide thickness on the tunnel current [15]. (a) Map of the tunnel current density  $J_G$  for fluctuations in the thickness  $e_{\text{ox}}$  with standard deviation  $\sigma_{\text{eff}} = 0.18$  nm. (b) Dependence of the total current density  $J_G$  on  $\sigma_{\text{eff}}$ . The simulation here concerns an NMOS transistor with gate length  $L_G = 70$  nm, width  $W = 200$  nm, and mean oxide thickness  $e_{\text{ox}} = 1.5$  nm. Fluctuations in  $e_{\text{ox}}$  are assumed to follow a Gaussian distribution. The bias on the transistor is such that  $V_{\text{GS}} = V_{\text{DD}} = 1$  V and  $V_{\text{DS}} = 0$  V, the state of the transistor for which the gate current is maximal

partly by the fact that, although  $I_{\text{on}}$  increases when  $L_G$  falls, the reduction in gate length tends to have the opposite effect on the tunnel current  $I_G$  since the area crossed by  $J_G$  becomes smaller. In addition, the reliability of the gate oxide seems to be enhanced, according to these studies, when the thickness of this layer becomes nanometric. Apparently, the breakdown field of the oxide layer would increase by 50% if  $e_{\text{ox}}$  were reduced from 5 nm to 1.5 nm. This improvement is probably due to the predominance of the direct tunnel effect, rather than other forms of tunneling effect, e.g., assisted by traps in the insulator, for oxide thicknesses below 2–3 nm and under low voltages ( $< 1.5$  V).

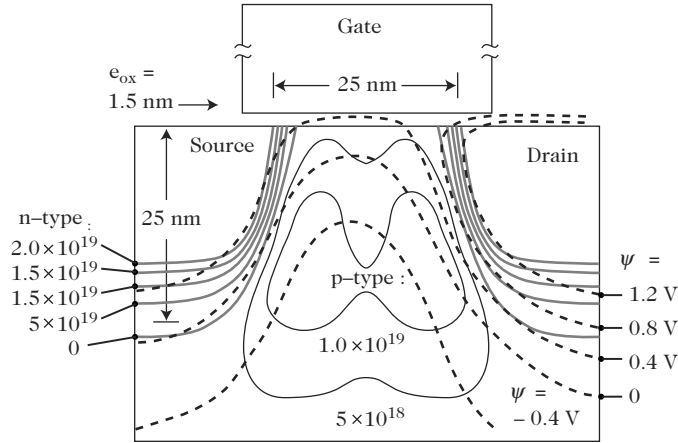
The increase in the gate leakage current is not only caused by the reduction in the oxide thickness, but at least as much by the inevitable fluctuations in this thickness which, poorly controlled, can give rise to ‘hot spots’ through which tunnel leakage is likely to be very high. This phenomenon is illustrated in Fig. 11.19, which shows the calculation results for the gate current, taking into account random fluctuations in the thickness according to a Gaussian distribution. For a nominal thickness of 1.5 nm and standard deviation of 0.2 nm for the fluctuations, it transpires that 80% goes through only 10% of the oxide surface area (see Fig. 11.19a). This kind of non-uniformity, comparable to the mean distance between atoms in Si (0.3 nm), increases the leakage current by a factor of 8 compared with an oxide layer with uniform nominal thickness (see Fig. 11.19b).

The encouraging results obtained for thicknesses in the vicinity of  $e_{\text{ox}} = 1.5$  nm, a value that was long considered to be unrealistic, illustrate the difficulty in defining an ultimate physical limit. This is exacerbated by the fact that such a limit may depend on the type of application envisaged. Today, it is considered that a thickness  $e_{\text{ox}}$  of the order of 1 nm may be feasible with an acceptable bound on the leakage current of 1–10 A/cm<sup>2</sup> for high performance technologies [16]. However, much lower leakage currents are required for low consumption applications. The results displayed in Figs. 11.18 and 11.19 show that this represents a genuine challenge.

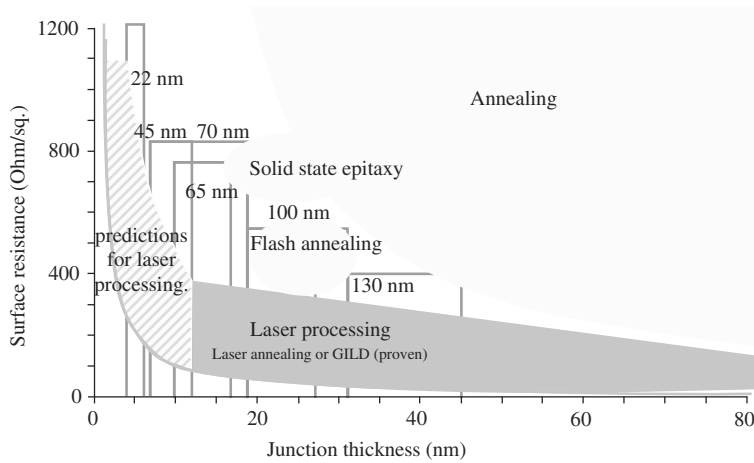
In order to take up this challenge, it would seem necessary, by around 2006–7, to replace the gate oxide by more electrostatically effective insulators, i.e., insulators with high relative dielectric permittivity  $\epsilon_r$ , known as high-k insulators. It would then be possible to increase the physical thickness in order to cut down the tunnel current, without reducing the control over the channel provided by the gate. Among the most promising high-k insulators, we may mention ZrO<sub>2</sub> and HfO<sub>2</sub>, with relative permittivities of the order of 25 in both cases, as compared with 3.8 for SiO<sub>2</sub>. Despite the intense research currently underway [8, 16, 17], the dielectric which will eventually replace SiO<sub>2</sub> is far from having been identified. One of the main difficulties is to obtain a high quality insulator/Si interface which does not degrade transport properties. There are also technological problems, such as thermodynamic stability on Si, film morphology, compatibility with the gate material and with a CMOS process in general, reliability, etc., and electrical problems due to the fact that the potential barrier at the high-k/Si interface is sometimes lower than that between SiO<sub>2</sub> and Si, a situation which favours the tunnel effect. Before moving on to genuine high-k insulators, nitrides ( $\epsilon_r = 7.9$ ) and oxynitrides are possible short term candidates [8]. These dielectrics have the advantage of limiting boron diffusion.

Another serious difficulty in the development of nanoMOS devices fulfilling the specifications is concerned with the problem of implementing the various doping requirements (channel and source/drain regions) designed to counter short channel effects. In the channel, highly retrograde profiles are needed, with low surface doping and ever higher buried doping in order to control punch-through. A highly 2D profile must also be obtained (halo or superhalo) (see Fig. 11.20), and this is very difficult to achieve in a repeatable manner [18]. Moreover, with the reduction in size of the active regions, the number (at most a few tens) and distribution of dopants become very difficult to control, whereas these very same parameters play a determining role in adjusting the electrical characteristics (threshold voltages) and performance of transistors. Their fluctuations from one transistor to another can make circuit operation highly problematic [19]. The only reasonable solution to this problem is probably to invent transistor architectures in which the channel is not doped, i.e., where doping is not used to shore up electrical characteristics and counter short channel effects. We shall return to this point in Sect. 11.3.2.



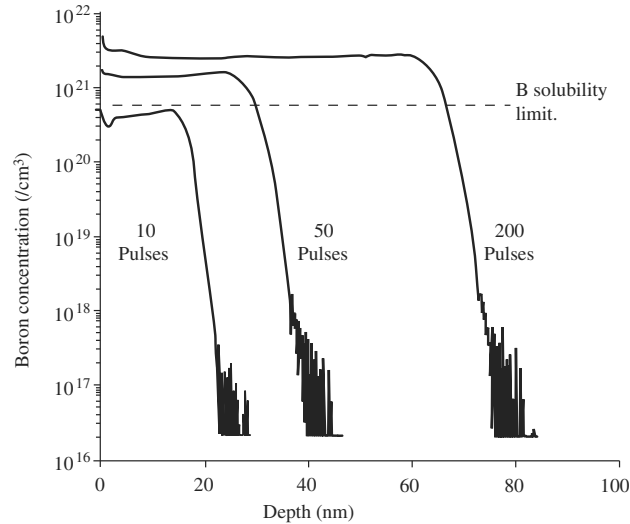


**Fig. 11.20.** NanoMOS architecture proposed by IBM with superhalo channel doping [20]. *Continuous curves* represent the boundaries of various doping regions, *n*-type for source and drain, and *p*-type for the channel. *Dashed curves* are loci of points with the same electric potential  $\psi$ , obtained by simulation



**Fig. 11.21.** Different techniques for obtaining ultrathin junctions in relation to the ITRS roadmap objectives (technological nodes 130, 100, . . . , 22 nm), illustrated by performance obtained in terms of junction depth  $X_j$  and surface resistance of doped layers, which conditions the value of the access resistance  $R_{acc}$  for entry into the channel in a MOSFET [21]

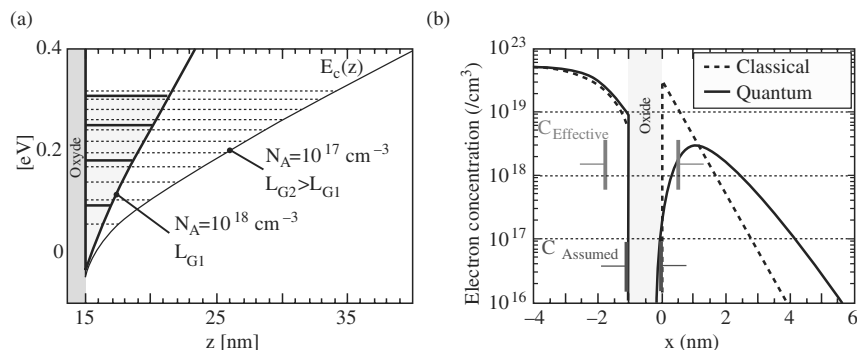
A similar problem is raised when doping the source and drain regions. As we saw in Sects. 11.2.3 and 11.2.4, the problem is to obtain junctions that are very shallow, once again to limit short channel effects, but also highly doped, and hence very abrupt, so that the contact resistance remains low compared



**Fig. 11.22.** SIMS (secondary ion mass spectroscopy) profiles of boron-doped samples obtained after 10, 50 and 200 sequences of gas injection followed by a laser pulse (GILD process). The *dashed line* indicates the boron solubility limit at thermal equilibrium, a maximum value that can be reached by doping methods used today [22]

with the channel resistance, which for its part tends to fall off when the transistor is made smaller. In this context, Fig. 11.21 clearly shows the limits of conventional doping techniques (implantation and annealing) for reconciling these aims. An alternative that is receiving more and more interest is based on the use of laser processes. One approach is laser thermal processing (LTP) after implantation, and another is gas immersion laser doping (GILD), a recently developed process which basically involves incorporating doping atoms contained in a gas, e.g.,  $\text{BCl}_3$ , into the Si by means of a laser pulse which makes the semiconductor surface amorphous [21, 22]. Indeed, LTP and GILD can produce high doping concentrations in small thicknesses (less than 20 nm), with very sharp profiles, as shown in Fig. 11.22. There is also a lesser penetration of the dopants under the oxide layer, whence a better control of the effective gate length. Recent progress obtained with excimer laser technology, especially in terms of energy, area and uniformity of the beam, should make it possible to process large areas and hence to use these techniques on the production line.

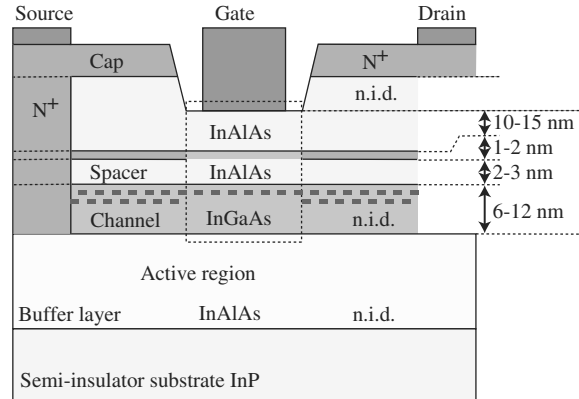
As we have already seen for the gate insulator with the tunnel current, the reduction in size down to nanometric values brings about new effects explicable only by quantum mechanics, which tend to modify the electrical characteristics of MOSFET devices. With the increase in channel doping in conventional CMOS technologies, the confining potential well in which the



**Fig. 11.23.** (a) Energy levels in the inversion layer of a MOS structure for two substrate doping levels ( $N_A = 10^{17} \text{ cm}^{-3}$  and  $10^{18} \text{ cm}^{-3}$ ) [23]. (b) Concentration profile in a MOS capacitor calculated with quantum confinement effects (*continuous curve*) and without (*dashed curve*). The  $n$ -doping of the polysilicon is  $5 \times 10^{19} \text{ cm}^{-3}$  and the  $p$ -doping of the silicon is  $10^{18} \text{ cm}^{-3}$ . The gate voltage is 0.3 V [24]

inversion layer forms (see Fig. 11.23) becomes ever narrower. The quantisation of energy levels then becomes more significant. In particular, the gaps between levels become larger, as shown in Fig. 11.23a. This phenomenon must be taken into account when scaling down nanoMOS devices. Indeed, due to the boundary conditions imposed on wave functions at the  $\text{SiO}_2/\text{Si}$  interface, the maximum carrier density is shifted by about 1 nm from the boundary with the oxide in the semiconductor (see Fig. 11.23b), a distance comparable with the thickness  $e_{\text{ox}}$ . If, moreover, we consider the surface depletion of the polysilicon layer forming the gate (also illustrated in Fig. 11.23b), we obtain an effective gate capacitance that is significantly lower than the theoretical value which does not take these effects into account. Overall, neglecting quantisation effects leads to an underestimate of the threshold voltage  $V_T$  (a typical shift would be of the order of 0.1 V) and an overestimate of the level of control by the gate [24]. Moreover, the depletion phenomenon in the gate polysilicon could be countered by using purely metal gates, which would also improved the frequency performance of MOS transistors due to the lower gate resistance.

Finally, the development of interconnect technologies to suit nanoMOS devices will probably be one of the most demanding tasks in the relatively short term. According to the 2001 ITRS roadmap, the total length of interconnects in a chip should be  $5 \text{ km/cm}^2$  in 2002,  $9 \text{ km/cm}^2$  in 2005, and  $11 \text{ km/cm}^2$  in 2007. As far as the last two figures are concerned, industrial solutions are known but have not yet been optimised. However, to achieve the value of  $16 \text{ km/cm}^2$  recommended for 2010 and  $22 \text{ km/cm}^2$  for 2013, industrial solutions are not yet known and remain to be developed if we hope to increase circuit operating frequencies as predicted (see Sect. 11.2.5). The development and integration of low- $k$  materials with ever smaller dielectric permittivities [25]



**Fig. 11.24.** HEMT architecture with double InAlAs/InGaAs heterojunctions on an InP substrate. n.i.d. means not intentionally doped

may not be able to meet the requirements, and it has been proposed to replace metal connections by electromagnetic or optical ones, which are much faster, at least for transmission between units and the distribution of the clock signal. In both cases, a great deal of design and technological work will be required to demonstrate that this is viable on a large scale whilst maintaining compatibility with front-end technologies used on the transistor level [26].

### 11.3.2 Alternatives to Conventional MOSFET Devices

The conventional MOSFET architecture illustrated schematically in Fig. 11.1 thus reaches its limits when the gate length goes below 50 nm, despite all the scaling rules and channel doping tricks (retrograde channel, pockets, halos) discussed in Sect. 11.2.3. To make nanoMOS devices that go beyond these limits, new transistor architectures must be devised.

The first new concept in this respect arose from band gap engineering techniques developed in the context of high speed components based on III–V semiconductors such as gallium arsenide GaAs. Band gap engineering is the practice of combining in the same device semiconductors with different values of the forbidden energy gap  $E_G$ . The difference in the value of  $E_G$  leads to discontinuities in the conduction band and or the valence band at the interface of these heterostructures. These discontinuities make it easier to juggle with the compromises required by scaling in the case of bipolar heterojunction transistors [27], but they are also used to produce high electron mobility transistors (HEMT) used in very high frequency units for telecommunications systems (see below).

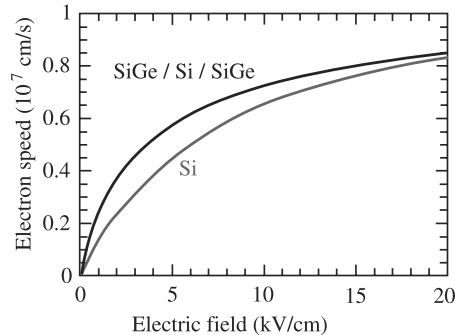
### III–V HEMT Devices

Figure 11.24 shows an HEMT transistor architecture based on III–V semiconductors. It uses a double InAlAs/InGaAs heterostructure controlled by a Schottky-type gate electrode and fabricated on a semi-insulating InP substrate. The upper InAlAs layer, with large band gap, is highly  $n$  doped over a thickness of 1–2 nm. This is known as  $\delta$ -doping. This region is fully depleted in terms of free carriers under the influence of the Schottky gate and carriers are transferred towards the InGaAs layer, the semiconductor of the heterostructure with small band gap. Depending on the value of the applied gate potential, a certain number of electrons will thus accumulate in a confinement well formed at the hetero-interface. In this way, the channel conduction can be modulated by the field effect as in a MOSFET.

The InAlAs layer separating the  $\delta$ -doping from the channel is not intentionally doped (n.i.d.), as is the InGaAs layer. This spacer serves to keep ionised impurities out of the conduction channel. The interface between the two semiconductors is generally very high quality. Finally, the mobility in InGaAs is intrinsically 10 times higher than in Si. All these factors mean that carrier mobility in the channel is much higher in HEMT devices than in Si MOSFET devices (see p.383). However, the former are only used in very high frequency applications, and ones which do not require a high integration density. This is explained by the higher cost of the basic materials and difficulties raised by the relevant technology. The development of III–V devices has also been slowed down by the fact that it is difficult to produce reliable insulators, and hence to build MOSFET-type architectures, with this type of semiconductor.

A form of band gap engineering can also be developed with Si, by using IV–IV alloys. For example, replacing a fraction  $x$  of the Si atoms in a monocrystalline structure by Ge atoms, one obtains an  $\text{Si}_{1-x}\text{Ge}_x$  type semiconductor with a smaller band gap than Si, but a larger crystal lattice parameter. Growing an SiGe layer on Si thus induces compressive mechanical stresses in the SiGe alloy. Conversely, a layer of Si deposited on an SiGe pseudo-substrate undergoes tensile stress. Heterostructures obtained in this way give rise to discontinuities in the conduction band and/or the valence band which can be exploited to confine electrons or holes [28]. Quite generally, the deformation of the electronic band structure caused by a compressive stress (SiGe on Si, or better still, Ge on SiGe) tends to enhance hole transport properties, whereas that caused by tensile stress (Si on SiGe) favours both hole and electron transport.

Si/SiGe heterostructures have already found an industrial application with the use of bipolar heterojunction transistors [27] in BiCMOS integrated circuits for telecommunications, such as GPS (Global Positioning System) receivers [29]. Thinking a little further ahead, HEMT devices based on SiGe/Si/SiGe heterostructures, therefore compatible with CMOS technology, have been under investigation now for several years. Very good performances have recently been reported at high frequencies for devices with gate length 100 nm [30]. Apart from the advantages for transport with the HEMT architecture (non-doped channel and good interface, see above), these results can be explained by the high mobility of electrons in the tensile-stressed Si channel



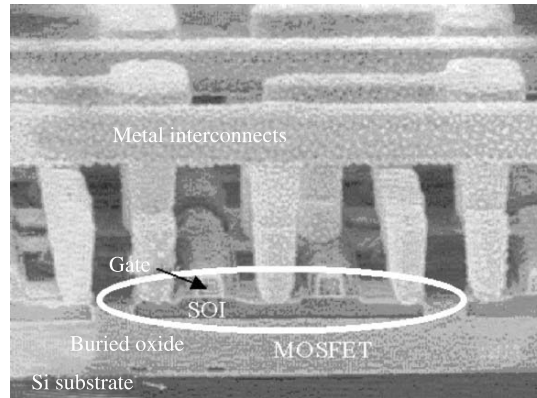
**Fig. 11.25.** Electron speed as a function of the electric field in the stationary transport regime at room temperature. Carriers are in the Si bulk or in an Si layer sandwiched between two SiGe layers. The mobility is given by the slope of the speed–field characteristic at the origin

between two SiGe layers, about twice as high as in bulk Si [31], as shown in Fig. 11.25. Likewise, compressive stress improves transport properties of holes confined in an SiGe layer sandwiched between two Si layers.

For CMOS technology, the beneficial effects of stresses on mobility can be exploited to increase the current  $I_{\text{on}}$ . The increased mobility of holes in a PMOS is also of great interest for the integration density (see p. 383). NMOS and PMOS transistors could then be based on double heterostructures SiGe/Si/SiGe and Si/SiGe/Si, respectively, with buried stressed channel. This configuration is also a solution for reducing the gate leakage current by the tunnel effect [32]. Furthermore, a simple Si/SiGe heterostructure might be envisaged for making the NMOS and PMOS with the same stack of layers in a surface channel architecture. Promising results have already been published with regard to the increase in mobility and in  $I_{\text{on}}$ , for both NMOS devices [33] and PMOS devices [34]. But the need to grow an SiGe pseudo-substrate a few microns thick on the Si substrate nevertheless makes integration of these devices somewhat problematic. Moreover, the fabrication of an oxide on SiGe, essential for the buried channel NMOS, has not yet been perfected.

Other possibilities are envisaged to avoid the use of a thick SiGe pseudo-substrate. For example, it is possible to transfer a thin layer of SiGe onto an oxide substrate and then deposit a tensile-stressed Si layer. Carbon can also be incorporated into the silicon to make an  $\text{Si}_{1-y}\text{C}_y$  or  $\text{Si}_{1-x-y}\text{Ge}_x\text{C}_y$  type stressed film (with  $y$  of the order of 1–2%) on an Si substrate, which could provide an alternative for NMOS fabrication. However, the incorporation of C in Si or SiGe is difficult [35] and the expected mobility enhancement in such carbon-bearing films has not yet been clearly demonstrated [36].

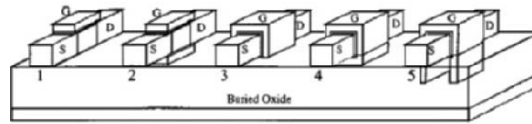
Although the use of mechanical stresses seems today to be a key line of inquiry for nanoMOS devices, to overcome the limitations of conventional structures as far as transport is concerned [37], it still does not really solve



**Fig. 11.26.** CMOS circuit using SOI technology [40]

the difficulties discussed earlier arising from short channel effects in such small devices. Now that the reduction in the thickness of the gate insulator is reaching its limits and channel doping techniques are becoming unrealistic, a major change in MOSFET architecture must be envisaged for the 2006–7 horizon. The very clear improvements brought by silicon-on-insulator (SOI) technology, in terms of both quality and cost, make it a very interesting prospect [38], and all the more so in that it remains compatible with current CMOS technologies and even with Si/SiGe heterostructures [39]. As can be seen from Fig. 11.26, SOI technology produces transistors on a thin layer of Si, with a thickness of a few tens of nanometers or less, separated from the substrate by a buried  $\text{SiO}_2$  layer, the buried oxide, often abbreviated to BOX. This architecture has an obvious advantage when it comes to limiting the bulk leakage current via the substrate (see the discussion of bulk punch-through in Sect. 11.2.2). If, moreover, the Si film is only slightly, or not at all doped, it may be wholly depleted of free carriers at zero applied gate voltage. This is also an advantage with regard to limiting the surface punch-through current. However, in this kind of architecture, the drain voltage can induce significant short channel effects due to electrostatic influences through the buried oxide, especially when the BOX is thick. The gate control of the Si film, and in particular at the interface with the buried oxide, will thus require some improvement.

In fact, the great contribution of SOI in improving the performance of nanoMOS devices can be put down to the possibility of building architectures (by transfer to other substrates, 3D structuring, etc.) in which the active zone of the transistor is commanded by the same gate voltage on two, three, or even four sides, as shown schematically in Fig. 11.27. The idea is that, if the thickness of the active Si film between the various gates is small enough, i.e., typically less than half the channel length in the case of a double gate structure (this criterion can be less restrictive with triple or quadruple gates),



**Fig. 11.27.** MOSFET on SOI with different gate architectures: (1) single gate, (2) double gate, (3) triple gate, (4) quadruple gate, and (5) proposed pi-shaped gate for better screening of the electrostatic influence of the drain through the buried oxide layer [41]

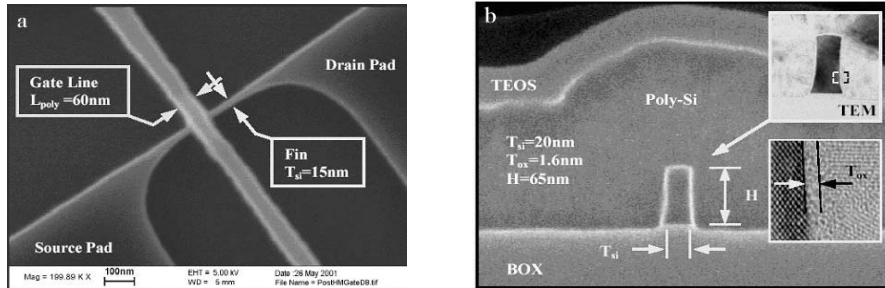
the gate voltage can command the whole volume of Si between the highly doped source and drain regions. For  $V_{GS}$  greater than the threshold voltage  $V_T$  (in absolute value), one then has a bulk rather than a surface inversion channel, for conducting the drain current. This is particularly favourable for the value of  $I_{on}$ . Moreover, the perturbing influence of  $V_{DS}$  on carrier injection from the source is then seriously limited, without the need to drastically reduce the gate insulator thickness or significantly increase channel doping levels. The active Si film can thus be unintentionally doped, which favours transport, although this means that a parameter other than the doping level must be found to adjust the value of  $V_T$ . The solution preferred at the moment to solve the latter problem consists in using a metal alloy for the gate. The composition of this alloy could then be adjusted to obtain a suitable threshold voltage. A metal gate would also provide a solution to problems raised by the polysilicon, as discussed earlier.

Transistors with multiple gate architecture are currently attracting much interest, in terms of both modelling and technology. Indeed, since the size of the active zone is reduced to a few tens of nanometers in all directions, the physics of transport in these devices raises a fair number of questions: the relevance of ballistic transport [42] or transport by the tunnel effect [43] between source and drain to name but two. Furthermore, nanoMOS fabrication involves rather delicate, almost acrobatic, processes. Figure 11.28 shows two views of the so-called FinFET architecture, where the word ‘fin’ refers to the shape of the active Si film. This structure exhibits very good electrical characteristics but requires a rather aggressive form of lithography [44].

## 11.4 Conclusion

The size reduction of MOSFET devices has led to a steady improvement in the performance of CMOS digital circuits, whilst controlling the relevant energy balances. However, this evolution cannot be pursued without observing a certain number of scaling rules which would appear to be approaching their limits as far as nanoMOS devices are concerned. The reduction in the gate





**Fig. 11.28.** FinFET architecture observed by electron microscopy [44]. (a) Top view. (b) Cross-sectional view perpendicular to the carrier flux between source and drain. The *inserts* are transmission electron microscopy (TEM) images of the fin and the gate oxide

oxide thickness down to about 1 nm thus raises serious difficulties of both a physical and an electrical order. In this context, it seems absolutely necessary to discover alternative architectures to those used in the ‘bulk’ MOSFET up to the present time. Multiple gate SOI devices, in which the size of the active zone decreases to a few tens of nanometers in all directions, and the use of stressed nanometric Si films should make it possible to improve the performance of CMOS circuits still further, at least as far as the 2012 horizon. In parallel with this evolution of standard microelectronics towards the ultimate electronics, preparatory studies must lay the foundations for alternatives to the CMOS logic.

## References

1. Moore, G.: *Electronics* **38** (8), 114–117 (1965)
2. Taur, Y., Ning, T.H.: *Fundamentals of Modern VLSI Devices*, Cambridge University Press (1998)
3. <http://www.microelectronique.univ-rennes1.fr/index21.html>
4. Borel, J.: Système sur une puce, RNRT Colloquium, Brest, France, 3–4 February 2002
5. Gwoziecki, R., Skotnicki, T., Bouillon, P., Gentil, P.: *IEEE Trans. Electron. Dev.* **46** (7), 1551–1561 (1999)
6. Thompson, S., Packan, P., Bohr, M.: *Intel Technology Journal* Q3’98 (1998)
7. <http://public.itrs.net>
8. Osburn, C.M., et al.: *IBM J. Res. Dev.* **46** (2/3), 299–315 (2002)
9. Deleonibus, S., et al.: *IEEE Electron. Dev. Lett.* **21** (4), 173–175 (2000)
10. Boeuf, F., et al.: *IEDM Tech. Digest* 637–640 (2001)
11. Chau, R., et al.: *IEDM Tech. Digest* 45–48 (2000)
12. Buchanan, D.A.: *IBM J. Res. Dev.* **43** (3), 245–264 (1999)
13. Rana, F., Tiwari, S., Buchanan, D.: *Appl. Phys. Lett.* **69** (8), 1104–1106 (1996)

14. Momose, H.S., Nakamura, S.-I., Ohguro, T., Yoshitomi, T., Morifuji, E., Morimoto, T., Katsumata, Y., Iwai, H.: IEEE Trans. Electron. Dev. ED **45** (3), 691–700 (1998)
15. Cassan, E., Dollfus, P., Galdin, S., Hesto, P., Microelectron. Reliab. **40** (4–5), 585–588 (2000)
16. Wilk, G.D., Wallace, R.M., Anthony, J.M.: J. Appl. Phys. **89** (10), 5243–5275 (2001)
17. Hiratani, M., Torii, K., Shimamoto, Y., Saito, S.-I.: Appl. Surf. Sci. **216** (1–4), 208–214 (2003)
18. ITRS Front End Processes, p. 28 (2001)
19. Asenov, A.: IEEE Trans. Electron. Dev. **45** (2), 2505–2513 (1998)
20. Taur, Y., Wann, C.H., Frank, D.J.: IEDM Tech. Digest 789–792 (1998)
21. Kerrien, G.: Doctoral thesis, Paris-Sud University XI, Orsay (2004)
22. Kerrien, G., Boulmer, J., Débarre, D., Bouchier, D., Grouillet, A., Lenoble, D.: Appl. Surf. Sci. **186** (1–4), 45–51 (2002)
23. Monsef, F.: Doctoral thesis, Paris-Sud University, Orsay (2002)
24. Cassan, E., Dollfus, P., Galdin, S., Hesto, P.: IEEE Trans. Electron. Dev. **48** (4), 715–721 (2001)
25. Maex, K., Baklanov, M.R., Shamiryan, D., Iacopi, F., Brongersma, S.H., Yanovitskaya, Z.S.: J. Appl. Phys. **93** (11), 8793–8841 (2003)
26. Laval, S.: C. R. Acad. Sci. Paris, t. 1, Series IV, 941–949 (2000)
27. Mouis, M.: Physique et technologie et environnement CMOS. In: *Physique des dispositifs pour circuits intégrés silicium* (EGEM treatise, series Electronique et micro-électronique), ed. by J. Gautier, Hermes Sciences-Lavoisier (2003) pp. 233–284
28. Galdin, S., Dollfus, P., Aubry-Fortuna, V., Hesto, P., Osten, H.J.: Semicond. Sci. Technol. **15** (16), 565–572 (2000)
29. Meyerson, B.S.: IBM J. Res. Develop. **44** (3), 391–407 (2000)
30. Aniel, F., Enciso-Aguilar, M., Giguere, L., Crozat, P., Adde, R., Mack, T., Seiler, U., Hackbarth, Th., Herzog, H.J., König, U., Raynor, B.: Solid-State Electron. **47** (2), 283–289 (2003)
31. Dollfus, P.: J. Appl. Phys. **82** (8), 3911–3916 (1997)
32. Cassan, E.: J. Appl. Phys. **87** (11), 7931–7939 (2000)
33. Jurczak, M., Skotnicki, T., Ricci, G., Campidelli, Y., Hernandez, C., Bensahel, D.: Proceedings of the 29th European Solid-State Device Research Conference, Louvain, Belgium (1999) pp. 304–307
34. Collaert, N., Verheyen, P., De Meyer, K., Loo, R., Caymax, M.: Solid-State Electron. **47** (7), 1173–1177 (2003)
35. Calmes, C., Bouchier, D., Débarre, D., Le Thanh, V., Clerc, C.: Thin Solid Films **428** (1–2), 150–155 (2003)
36. Osten, H.J., Gaworzewski, P.: J. Appl. Phys. **82** (10), 4977–4981 (1997)
37. Thompson, S., et al.: IEDM Tech. Digest 61–64 (2002)
38. Allibert, F., Ernst, T., Pretet, J., Hefyene, N., Perret, C., Zaslavsky, A., Cristovaleanu, S.: Solid-State Electron. **45** (4), 559–566 (2001)
39. Cheng, Z.-Y., Currie, M.T., Leitz, C.W., Taraschi, G., Fitzgerald, E.A., Hoyt, J.L., Antoniadis, D.A.: IEEE Electron. Dev. Lett. **22** (7), 321–323 (2001)
40. <http://www-3.ibm.com/chips/gallery/soichip.html>
41. Park, J.-T., Colinge, J.-P., Diaz, C.H.: IEEE Electron. Dev. Lett. **22** (8), 405–406 (2001)

42. Rhew, J.-H., Ren, Z., Lundstrom, M.S.: *Solid-State Electron.* **46** (11), 1899–1906 (2002)
43. Mouis, M., Poncet, A.: ESSDERC (European Solid-State Device Research Conference), Nuremberg (Germany), 11–13 September 2001, pp. 211–214
44. Kedzierski, J., et al.: *IEDM Tech. Digest* 437–440 (2001)

## Alternative Electronics

J.-N. Patillon and D. Maily

As discussed in previous chapters, the introduction of new technologies has made it possible, and will continue to make it possible, to extend CMOS electronic components down towards nanoscale MOSFET structures. A cautious optimist might expect these technologies to perpetuate the reign of the MOSFET as far ahead as 2016.

Beyond the CMOS, completely new approaches are beginning to emerge, for both data processing and storage. The aim of the present chapter is to review the physical ideas underlying these novel families of components, born from the current and future requirements of nanoscience.

In this chapter, three categories of components will be discussed, defined by the type of physical properties they exploit:

- The first category comprises single-electron devices (SED), based on the controllable transfer of individual electrons between conducting islands. Recent research in this area has already generated several ideas for components which could revolutionise the field of memory cells and digital data storage.
- The second category collects together mesoscopic components, based on the prospects offered by nanotechnology for fabricating small structures. Indeed, such structures open the way to a new regime of mesoscopic devices. In this regime, the phase coherence of the electron wave function and quantum interference modify ordinary Boltzmann transport, whilst electron correlations can alter the behaviour of bulk materials in confined structures.
- The last category is based on the use of superconducting materials rather than semiconductors. These new components are very different from those developed in the USA at the beginning of the 1980s. The new approach, known as rapid single-flux quantum (RSFQ) logic, is based on the storage and transmission of magnetic flux quanta.

All these components share one common factor: they exploit physical properties due to the size of the materials used. It is the use of these properties

which has allowed research scientists to create novel functions with the help of such components.

## 12.1 Characteristic Length Scales for Nanoscopic Components

In the ohmic regime, the conductance  $G$  of a 3D rectangular conductor is directly proportional to the cross-sectional area  $S$  and inversely proportional to its length  $L$ , i.e.,

$$G = \sigma \frac{S}{L}, \quad (12.1)$$

where the conductivity  $\sigma$  is constant throughout the conductor. The relentless quest to reduce the size of integrated circuits observed in microelectronics raises the question as to how far the size can be reduced before this ohmic behaviour will disappear. Beyond this point, a new category of systems comes into being: nanoscopic systems, which are bigger than the atomic scale but not big enough to be ohmic (see Chap. 11).

There are three characteristic lengths [1]:

1. The Fermi wavelength,

$$\lambda_F = \frac{2\pi}{k_F}, \quad (12.2)$$

related to the electron Fermi energy

$$\varepsilon_F = \frac{\hbar^2 k_F^2}{2m^*}. \quad (12.3)$$

In fact,  $\lambda_F \approx 10$  nm for semiconductors and  $\lambda_F \approx 1$  nm for metals.

2. The mean free path  $l$ , which is the distance travelled by an electron before its initial momentum is destroyed. For a high quality semiconductor,  $l \approx 1$   $\mu\text{m}$ .
3. The phase relaxation length or coherence length  $L_\varphi$ , which is the distance travelled by an electron before the initial phase of its wave function is destroyed ( $L_\varphi \approx 1$   $\mu\text{m}$ ), longer than the mean free path.

These quantities vary from one material to another and are significantly affected by temperature, magnetic field, etc. A conductor will have ohmic behaviour if its dimensions are much bigger than each of these characteristic lengths.

When one seeks to describe transport properties in structures with dimensions of the same order of magnitude as the de Broglie wavelength, the nonlocality of particles rules out the simplifying approximation whereby collisions are treated as instantaneous and independent of the phase. It is this

approximation that allows one to decouple the higher orders of the distribution functions in the Boltzmann equation.

For structures with characteristic lengths  $L$  smaller than the mean free path, i.e.,  $L \ll l$ , it may be assumed that particles move through the active region without scattering, giving rise to what is known as ballistic transport. If the total length of the system is smaller than the coherence length, i.e.,  $L \ll L_\varphi$ , phase coherence can be maintained over a long enough distance to justify describing electrons with a wave function that extends over the whole system.

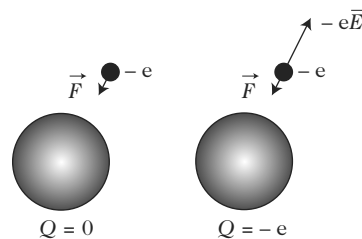
In conclusion, nanoscopic components of type SED, MC or RFSQ are components for which the characteristic lengths of the active region are of the same order of magnitude as or smaller than the Fermi wavelength. Charge carriers can then be described using a quantum formalism and we shall see in the following how the ensuing quantum effects can be exploited to create novel functions.

## 12.2 Single-Electron Devices (SED)

### 12.2.1 Basic Ideas

Electrons were first manipulated on an individual basis in Millikan's experiments at the beginning of the twentieth century. However, their use in components only became possible in the last decade or so, for this requires nanoscopic structures which could not be fabricated until recently. These new nanofabrication tools, like the novel structures (e.g., single-molecule structures) which have appeared over the last ten years, have opened the way to single-electron electronics. The underlying idea is the phenomenon of Coulomb blockade. Before outlining this phenomenon, let us discuss the physical principles that govern it.

Consider a small conductor, referred to conventionally as an island, which starts with zero charge, i.e., it contains exactly the same number  $m$  of electrons and protons (see Fig. 12.1). In this state, the island generates no electric field outside of itself. A weak force  $\vec{F}$  is therefore sufficient to bring an electron



**Fig. 12.1.** Neutral island acquiring an electron

to the island from elsewhere in the crystal. The net charge on the island is then  $-e$  and the resulting electric field  $\mathbf{E}$  repels other electrons in the neighbourhood. Although the electron charge is very small,  $e \approx 1.6 \times 10^{-19}$  C, the electric field  $\mathbf{E}$  at the surface of the island is inversely proportional to the square of the size of the island and it soon becomes very large in the case of nanostructures, e.g., 140 kV/cm at the surface of a sphere of diameter 10 nm.

In fact, the electric field no longer provides an adequate measure of these effects. A better measure is the electrostatic energy arising from the fact that each additional charge  $dq$  brought to the island must dissipate an energy

$$E_C = \frac{q^2}{2C} \quad (12.4)$$

against the field due to the charges already present in the island. When the size of the island is comparable with the de Broglie wavelength, the quantisation energy of the electron

$$E_a = \frac{1}{2m^*} \left( \frac{\hbar\pi N}{d} \right)^2, \quad (12.5)$$

where  $m^*$  is the effective mass,  $d$  the island diameter, and  $N$  the electron density, becomes non-negligible and has to be added in. For islands with very small sizes, of the order of the nanometer, the last formula is no longer valid because the notion of effective mass is based on the period of the crystal lattice. However, it provides an acceptable approximation for studying the related physical effects.

### 12.2.2 Transport by Coulomb Blockade

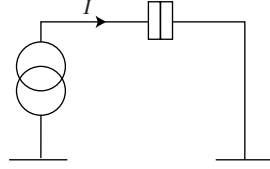
#### Tunnel Effect and Tunnel Junction

The tunnel effect is a phenomenon predicted by quantum mechanics, which attributes to particles, and in particular to electrons, a nonzero probability of crossing a potential barrier with a behaviour described by a wave function obeying the Schrödinger equation (see Chap. 3). It is thus possible to cross a potential barrier even when the particle has lower energy than the barrier height. The transparency of the barrier can be calculated from the Schrödinger equation and it is characterised by the transmission coefficient  $T$  expressed by

$$T = \frac{(2k\kappa)^2}{(k^2 + \kappa^2)^2 \sinh^2(2a\kappa) + (2k\kappa)^2}, \quad (12.6)$$

where

$$k = \sqrt{\frac{2mE}{\hbar^2}}, \quad \kappa = \sqrt{\frac{2m(V_0 - E)}{\hbar^2}},$$



**Fig. 12.2.** Tunnel junction

and  $2a$  is the width of the potential barrier.

The basic configuration of any single-electron device is a potential barrier whose geometric shape allows an electron to transfer by the tunnel effect between two conductors. Such a setup is usually called a tunnel junction.

This junction can be characterised by elements from conventional circuit theory, as shown in Fig. 12.2: capacitance  $C$  and resistance  $R$ . The resistance  $R$  is related to the opacity of the barrier, whilst  $C$  represents the capacitance of the insulator (the barrier) between two (semi-) conducting elements. This capacitance represents the fact that charges can accumulate against the barrier.

### Coulomb Oscillations

If the electron current in a conductor is interrupted by a tunnel junction, it is discretised since the electrons accumulate behind the barrier until an electron acquires an energy allowing its transfer through this barrier. If the junction is supplied with a current of intensity  $I$ , this current will then oscillate in time. Such oscillations, known as Coulomb oscillations, can be explained by the basic principles of thermodynamics, i.e., by the minimisation of the free energy defined by the difference between the total energy stored in the system and the work done by the voltage supply:

$$\Delta F = \Delta E_{\Sigma} - \Delta W, \quad (12.7)$$

where  $E_{\Sigma} = E_C + \Delta E_F + E_N$ , and  $\Delta E_F$  represents the variation in the Fermi level. To a first approximation, the last two terms (Fermi energy and quantum energy) can be neglected.

### Characteristic Energies

The electrostatic energy of the charge obtained when a condenser is charged up between two conductors is

$$\Delta E_C = \frac{Q^2}{2C}.$$

The work done by voltage sources is



$$\Delta W = \sum_{\text{sources}} \int V(t)I(t)dt .$$

For transfer by the tunnel effect, this implies a contribution  $eV$  due to the transferred electron and a contribution due to polarisation charges created in response to the change in potential of the island. For constant voltage sources  $V_i$ , the total work is given by

$$\Delta V = \pm eV + \sum_{\text{sources}} V \Delta q .$$

The quantum confinement energy is

$$E_N = \frac{\hbar^2}{2m^*} \left( \frac{\pi N}{d} \right)^2 ,$$

where  $m^*$  is the effective electron mass and  $d$  the depth of the potential well.

### Orthodox Theory

Returning to the island discussed above to introduce the basic principle of single-electron devices and defining it as the connection in series of two tunnel junctions as shown in Fig. 12.3, we shall now show how the Coulomb blockade effect can be explained in a simple manner by the orthodox theory proposed by D. Averin and K. Likharev. This theory is based upon the following simplifying hypotheses:

- The quantisation energy of the electrons is neglected, i.e., the energy spectrum of the electrons is treated as continuous. This assumption is only valid if  $E_k \ll k_B T$ , but it provides an adequate description whenever  $E_k \ll E_C$ .
- The transfer of an electron across the barrier by the tunnel effect is instantaneous.
- Coherent quantum processes involving simultaneous tunnel events are ignored. This assumption is valid if the resistance  $R$  of all tunnel barriers is much greater than the quantum resistance  $R_Q$ , i.e.,  $R \gg R_Q$ , where  $R_Q = h/4e^2 \approx 6.5 \text{ k}\Omega$ .

The rate of tunnel events between an initial state labelled  $i$  and a final state labelled  $f$  is given by the Fermi golden rule as

$$\Gamma_{i \rightarrow f}(\Delta F) = \frac{2\pi}{\hbar} |T_{k_i, k_f}|^2 \delta(E_i - E_f - \Delta F) , \quad \Delta F = F_f - F_i . \quad (12.8)$$

The difference between the initial and final energies of the tunnel electron,  $E_i$  and  $E_f$ , respectively, includes the free energy variation it generates. The total rate of tunnel events from occupied states on one side to unoccupied states on the other is given by the overall expression

$$\Gamma(\Delta F) = \frac{2\pi}{\hbar} \sum_i \sum_f |T_{k_i, k_f}|^2 f(E_i) [1 - f(E_f)] \delta(E_i - E_f - \Delta F) , \quad (12.9)$$

where  $f(E)$  is the Fermi–Dirac distribution giving the occupation probability of the levels:

$$f(E) = \frac{1}{1 + \exp \frac{E - E_F}{k_B T}} .$$

### Fermi Golden Rule

The time-dependent Schrödinger equation can be written in the form

$$(H_0 + H_1)\psi = i\hbar \frac{\partial \psi}{\partial t} ,$$

where  $H_1$  is a small perturbation of the Hamiltonian  $H_0$ . The solution in the unperturbed case is given by  $H_0\psi_n = \hbar\omega_n\psi_n$ , where  $\int |\psi_n|^2 d^3r = 1$ , for  $n = 1, 2, \dots$ . Assuming that the perturbation remains small, the wave function in a standard treatment can be expanded in a series of orthonormal eigenfunctions of the unperturbed Hamiltonian:

$$\psi = \sum_n c_n(t)\psi_n e^{-i\omega_n t} .$$

Inserting this in the time-dependent Schrödinger equation and multiplying by  $\psi_m^* e^{i\omega_m t}$ , we obtain

$$i\hbar \frac{dc_m(t)}{dt} = \sum_n c_n(t) \int \psi_m^* H_1 \psi_n e^{it(\omega_m - \omega_n)} .$$

Setting  $\omega_{mn} = \omega_m - \omega_n$  and  $H_{mn} = \int \psi_m^* H_1 \psi_n d^3r$ , the last equation becomes

$$i\hbar \frac{dc_m(t)}{dt} = \sum_n c_n(t) H_{mn} e^{i\omega_{mn} t} .$$

We now make the following assumptions:

1. The perturbation begins at time  $t = 0$  and hence  $c_m(0) = 1$  and  $c_n(0) = 0$  for all the other states.
2. Scattering out of the initial state is low, so that  $c_m(t) = 1$ , an assumption which neglects the principle of conservation of particle number.

We then deduce that

$$c_n(t) = -\frac{i}{\hbar} \int_0^t H_{mn} e^{i\omega_{mn} t} = \frac{H_{mn}}{\hbar\omega_{nm}} (1 - e^{i\omega_{nm} t}) .$$

The occupation probability of state  $n$  can thus be written

$$|c_n(t)|^2 = \frac{4|H_{mn}|^2 t^2 \sin\left(\frac{\omega_{mn} t}{2}\right)}{\left(\frac{\omega_{mn} t}{2}\right)^2} .$$

When  $t$  is large enough to be sure that the scattering process has ended, the function

$$\frac{\sin^2\left(\frac{\omega_{mn}t}{2}\right)}{\left(\frac{\omega_{mn}t}{2}\right)^2}$$

tends to a Dirac  $\delta$ -function.

The transition probability per unit time or transmission rate is then given by the Fermi golden rule

$$\Gamma = \frac{d|c_n(t)|^2}{dt} = \frac{2\pi}{\hbar} |H_{nm}|^2 \delta(E_m - E_n) ,$$

which indicates that, in our case of tunnel processes, scattering only occurs if the particle energy is conserved.

A reasonable approximation can be made by neglecting the variation of the tunnel transmission coefficient with the energy. The transmission probability  $|T|^2$  can then be treated as a constant and taken outside the sum to give

$$\Gamma(\Delta F) = \frac{2\pi}{\hbar} |T|^2 \sum_i \sum_j f(E_i) [1 - f(E_f)] \delta(E_i - E_f - \Delta F) . \quad (12.10)$$

The number of electrons in an energy interval  $dE$  is given by  $D(E)dE$ , where  $D(E)$  is the density of states. We may then convert the sums into integrals, whence

$$\Gamma(\Delta F) = \frac{2\pi}{\hbar} |T|^2 \int_{E_{c,i}}^{\infty} dE_i \int_{E_{c,f}}^{\infty} dE_f D(E_i) D(E_f) f(E_i) [1 - f(E_f)] \delta(E_i - E_f - \Delta F) , \quad (12.11)$$

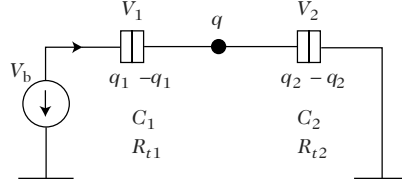
where  $E_{c,\{i,f\}}$  is the conduction band edge from which the electrons tunnel ( $i$ ) and towards which they tunnel ( $f$ ). Likewise  $D_{\{i,f\}}(E_{\{i,f\}})$  is the density of states of the initial and final barriers. The product  $f(E_i)[1 - f(E_f)]$  defines a quasi-rectangular function between the energies  $E_i$  and  $E_f$ . Since the contribution of the integral comes from this rectangular function, the density of states can be considered as constant and brought outside the integral. The delta function can be used to carry out one of the integrations in (12.11), which yields

$$\Gamma(\Delta F) = \frac{2\pi}{\hbar} |T|^2 D_i D_f \int f(E) [1 - f(E - \Delta F)] dE , \quad (12.12)$$

where  $E_c = \max(E_{c,i}, E_{c,f})$ .

Introducing the tunnel resistance, which incorporates the transmission probability and the densities of states,

$$R_T = \frac{\hbar}{2\pi e^2 |T|^2 D_i D_f} , \quad (12.13)$$



**Fig. 12.3.** Equivalent circuit for a double tunnel junction

and carrying out the integration, we obtain the main result of the orthodox theory of the single-electron tunnel effect:

$$\Gamma(\Delta F) = -\frac{\Delta F}{e^2 R_T (1 - e^{\Delta F/k_B T})}. \quad (12.14)$$

This shows that the probability of an electron transfer entailing an increase in the Helmholtz energy (energetically unfavourable) is very low. At absolute zero temperature, it reduces to

$$\Gamma(\Delta F) = \begin{cases} 0, & \Delta F \geq 0, \\ \frac{-\Delta F}{e^2 R_T}, & \Delta F < 0. \end{cases} \quad (12.15)$$

As mentioned at the beginning of the chapter, the Coulomb blockade model is only valid under certain conditions, in particular, if the electron states are localised in the island.

If we consider the characteristic fluctuation time  $\Delta t \approx R_T C$  of the charge and the energy jump  $\Delta E = e^2/C$  for an electron, then the uncertainty relation  $\Delta E \Delta t > h$  implies the following condition for the existence of a tunnel resistance:

$$R_T > \frac{h}{e^2} = R_Q = 25\,813 \, \Omega.$$

Owing to the restrictive hypotheses formulated above, the orthodox theory does not take into account a certain number of physical effects such as, for non-superconducting systems, cotunneling (the simultaneous occurrence of several tunnel events) or effects due to the discretisation of energy levels [2].

### 12.2.3 Double Tunnel Junction

Let us return to the island comprising two tunnel junctions connected in series, associating the values  $C_1, R_1$  and  $C_2, R_2$  in the obvious way. If the initial charge in the island is  $q_0$ , the charges in junctions 1 and 2 and the total charge can be written  $q_1 = C_1 V_1 = n_1 e$ ,  $q_2 = C_2 V_2 = n_2 e$  and  $q = q_2 - q_1 + q_0 = -n e + q_0$ , where  $n_1$  and  $n_2$  are the numbers of electrons

tunneling through junctions 1 and 2, respectively, and  $n = n_1 - n_2$  is the total number of electrons in the island.

The charge  $q_0$  is generally a non-integer charge shift, being due to parasitic capacitances or impurities located near the island, always present in practice. If  $V_b$  is the sum of the voltages across each of the junctions, i.e.,  $V_b = V_1 + V_2$ , we soon find that

$$V_1 = \frac{C_2 V_b + ne}{C_\Sigma}, \quad V_2 = \frac{C_1 V_b - ne}{C_\Sigma}, \quad C_\Sigma = C_1 + C_2. \quad (12.16)$$

We thus obtain the electrostatic energy stored in the double junction as

$$E_C = \frac{q_1^2}{2C_1} + \frac{q_2^2}{2C_2} = \frac{C_1 C_2 V_b^2 + (ne)^2}{2C_\Sigma}. \quad (12.17)$$

Finally, we may calculate the free energy by considering the work supplied by the voltage supply. If an electron tunnels through the first junction, the voltage supply must replace this electron ( $-e$ ) plus the change in voltage induced by the tunnel event. The voltage  $V_1$  changes by the amount  $-e/C_\Sigma$  and the charge changes by  $-eC_1/C_\Sigma$ . The charge  $q_1$  is reduced, indicating that the voltage supply 'receives' this charge. The total charge that must be replaced by the source is then  $-eC_2/C_\Sigma$  and the work done by the source in the case of tunnel events through both junctions 1 and 2 is

$$W_1 = -\frac{n_1 e C_2}{C_\Sigma} V_b, \quad W_2 = -\frac{n_2 e C_1}{C_\Sigma}. \quad (12.18)$$

The complete free energy for the circuit is then given by

$$F(n_1, n_2) = E_C - W = \frac{1}{C_\Sigma} \frac{1}{2} \left[ C_1 C_2 V_b^2 + (ne)^2 + e V_b (C_1 n_2 + C_2 n_1) \right]. \quad (12.19)$$

The changes in free energy for an electron which tunnels through junction 1 and junction 2 are given respectively by

$$\Delta F_1^\pm = F(n_1 \pm 1, n_2) - F(n_1, n_2) = \frac{e}{C_\Sigma} \left[ \frac{e}{2} \pm (V_b C_2 + ne) \right] \quad (12.20)$$

and

$$\Delta F_2^\pm = F(n_1, n_2 \pm 1) - F(n_1, n_2) = \frac{e}{C_\Sigma} \left[ \frac{e}{2} \pm (V_b C_1 - ne) \right]. \quad (12.21)$$

The probability of a tunnel event will only be large if the change in free energy is negative, indicating a transition to a state of lower energy. This is a direct consequence of the orthodox theory. It follows that  $\Delta F$  will be positive as long as the bias  $V_b$  is greater than a threshold which is a function of the smaller of the two capacitances. For a symmetric junction, i.e.,  $C_1 = C_2 = C_\sigma/2$ , this condition becomes  $|V_b| > e/C_\Sigma$ .

Thermodynamics thus provides a way of understanding this suppression of the tunnel effect, which causes the charge blockage known as Coulomb blockade. We also see from (12.20) and (12.21) that this phenomenon results from the electrostatic energy expended to allow an electron to enter or leave the island.

When the tunnel transition condition is fulfilled and an electron has been transferred to the island, in junction 1 for example, the island will possess  $n = 1$  excess electrons. According to what was said above, it is then energetically favourable for an electron to leave the island via the other junction and no other electron will be able to cross the first junction until this event has occurred. We thus observe a correlated set of electron transfers which gives rise to a resistive current of resistance  $R_t = R_{t1} = R_{t2}$  in the case of a symmetric junction.

On the other hand, it is clear that an asymmetric junction, e.g.,  $R_{t2} \gg R_{t1}$  will favour the accumulation of electrons in the island since, for an electron to leave it, a high bias will be needed to allow the immediate transfer of another electron through the first junction. The island will thus remain populated. The number of electrons in the island will increase for higher voltages, thereby creating a stair-shaped charge–voltage characteristic.

### 12.2.4 Single-Electron Transistor

Finally, adjoining a gate electrode to this double tunnel junction, the current flow through the island can be controlled by capacitive coupling. We have then created a single-electron transistor (SET) with structure given schematically by the equivalent circuit diagram in Fig. 12.4.

The effect of the gate electrode, represented by the capacitance  $C_g$ , is to modify the charge in the island, because the gate polarises the island. The total charge then becomes

$$q = -ne + C_g(V_g - V_2) . \tag{12.22}$$

This induces a change in the voltages  $V_1$  and  $V_2$  by addition of the charge  $C_g(V_g - V_2)$ . The voltages across the two junctions then become

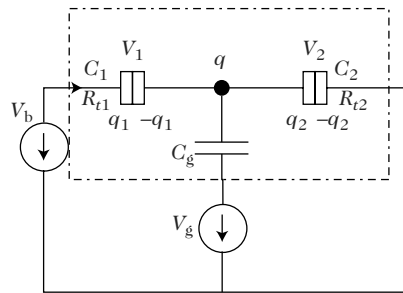


Fig. 12.4. Single-electron transistor

$$V_1 = \frac{(C_2 + C_g)V_b - C_g V_g + ne}{C_\Sigma}, \quad V_2 = \frac{C_1 V_b + C_g V_g - ne}{C_\Sigma}, \quad (12.23)$$

where  $C_\Sigma = C_1 + C_2 + C_g$ .

As a consequence, the electrostatic energy also includes the energy stored in the gate capacitance and the work done by the gate changes the free energy. This induces a change in the free energies of the two junctions after a tunnel event:

$$\Delta F_1^\pm = \frac{e}{C_\Sigma} \left\{ \frac{e}{2} \pm [(C_2 + C_g)V_b - C_g V_g + ne] \right\} \quad (12.24)$$

and

$$\Delta F_2^\pm = \frac{e}{2} \left[ \frac{e}{2} \pm (C_1 V_b + C_g V_g - ne) \right]. \quad (12.25)$$

### Stability Diagram

We shall now discuss an intuitive representation of the physical behaviour involved here. The first step is the construction of a stability diagram. In order to construct this, we first consider the case where the change in free energy is zero:

$$\left. \begin{array}{l} \Delta F_1^\pm = 0 \\ \Delta F_2^\pm = 0 \end{array} \right\} \implies \left\{ \begin{array}{l} \frac{e}{2} \pm [(C_2 + C_g)V_b - C_g V_g + ne] = 0, \\ \frac{e}{2} \pm (C_1 V_b + C_g V_g - ne) = 0. \end{array} \right. \quad (12.26)$$

This situation can be depicted graphically by expressing  $V_b$  as a function of  $V_g$  for fixed values of the capacitances and resistances. Since the two relations depend on the number  $n$  of electrons in the island, the graph takes the form of a series of parallel straight lines for each of the two expressions. In these expressions, the slopes of  $V_b$  as a function of  $V_g$  have opposite signs, so that the two groups of straight lines intersect as shown in Fig. 12.5.

The shaded areas correspond to stable regions with a whole number of excess electrons in the island. The blockade zones then appear as parallelograms. If the gate bias is increased, the drain-source bias  $V_b$  remains fixed below the Coulomb blockade voltage. This is equivalent to cutting the stability region parallel to the horizontal axis with a period  $e/C_g$ . This induces a current with Coulomb oscillations whose period is a function of the applied bias, where regions without tunneling alternate with regions of correlated tunnel events. The bigger the value of  $V_b$  and the more closely it approaches the Coulomb blockade voltage, the greater the width of the oscillations.

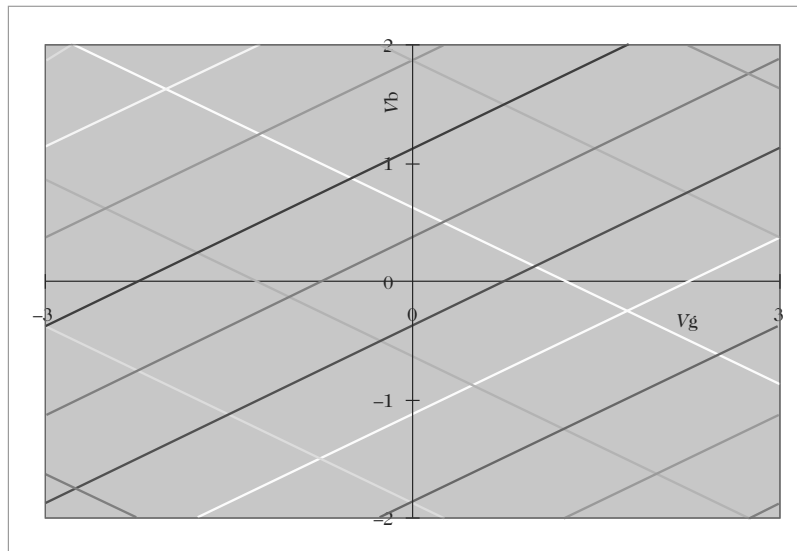
## 12.3 Quantum Interference in Nanostructures

### 12.3.1 Introduction

The wave nature of electrons is a crucial feature of atomic physics. In solid state physics, it is the Bragg diffraction of electron wave functions on crystal

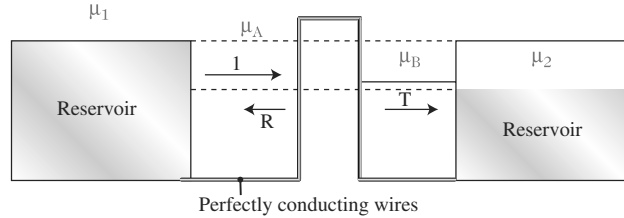
planes which leads to the appearance of energy bands. However, the electron transport properties of solids are well described by theories that make no mention of quantum interference. For example, the Drude transport theory, which provides good predictions concerning the conductivity of a great many metals and allows a deduction of the Wiedemann–Franz relation, treats electrons in a metal as a gas of classical particles. Naturally, a more precise description requires an application of the Pauli principle (the Bohr–Sommerfeld theory), and interactions between particles must be taken into account (Hartree and Hartree–Fock theories). But none of these theories involve the wave nature of electron wave functions.

However, with the advent of nanotechnology, it has become possible recently to carry out measurements on very small samples which have shown clear deviations from the usual results obtained with macroscopic samples. Ideas as firmly established as the additivity of series resistances are no longer respected in such samples. Such discrepancies are caused by interference between the different electron paths during propagation through the sample. For this interference to occur, phase coherence must be preserved throughout the sample. One thus introduces the notion of phase coherence length, the distance over which the phase of the electron wave function is conserved. The electron may undergo a great many elastic collisions which modify the phase of the wave function but in a perfectly well defined manner. However, any inelastic collision, exchanging energy with the external medium, tends to scramble this phase coherence, destroying the interference pattern and restoring the ‘classical’ behaviour of a macroscopic solid. In pure metals, the phase



**Fig. 12.5.** Stability diagram





**Fig. 12.6.** Model for a sample disordered by a barrier with transmission  $T$

coherence length  $L_\varphi$  may be as long as a few microns at very low temperatures. A mesoscopic system, i.e., of intermediate size, is a sample that is much bigger than a microscopic system, but which nevertheless requires a quantum description.

### 12.3.2 Conductance and Transmission. The Landauer Formula

Resistance is naturally associated with a notion of dissipation. In mesoscopic systems, only elastic scattering can occur. One must therefore reassess the idea of resistance in these samples without dissipation. It was Landauer who first carried out this reformulation of the idea of resistance. He modelled a sample by a barrier with transmission  $T$  and reflection  $R$  (see Fig. 12.6). To carry out measurements on the sample, reservoirs are attached via perfect conductors. The role of these reservoirs is to inject electrons in a state defined by their electrochemical potential  $\mu$  and to absorb electrons which reach them. It is thus in the reservoirs that energy dissipation occurs. The perfectly conducting wires transmit electrons adiabatically into the sample. In order to impose a current through the system, a potential difference is applied across the reservoir terminals. It must be small enough to ensure that neither the transmission nor the density of states varies over this energy range.

Consider first the 1D case, where there is only one channel. By the conduction channel, we understand the mode of propagation characterised for example by the quantisation of the transverse wave vector. Typically, for a sample with cross-sectional area  $A$ , there are  $N = A/\lambda_F^2$  conduction channels, where  $\lambda_F$  is the Fermi wavelength.

Refer now to Fig. 12.6. Since below  $\mu_2$  all states are occupied on either side of the reservoirs, there are as many electrons going from right to left as there are going from left to right. The overall current is therefore zero for these energies. We shall thus be concerned with the band between  $\mu_2$  and  $\mu_1$ . There are a total of  $2(\mu_1 - \mu_2)\partial n/\partial E$  states available in this band, where  $\partial n/\partial E$  is the density of states for just one direction. The chemical potentials  $\mu_A$  and  $\mu_B$  in the wires are determined by the equilibrium between the number of electrons above and the number of free states (holes) below.

The total current between the reservoirs is then given by

$$ev(\mu_1 - \mu_2)T \frac{\partial n}{\partial E},$$

where  $v$  is the speed of particles at the Fermi level. Since in 1D,

$$\frac{\partial n}{\partial k} = \frac{\partial n}{\partial E} \frac{\partial E}{\partial k} = \frac{1}{\pi},$$

it follows that

$$\frac{\partial n}{\partial E} = \frac{2}{hv},$$

and the current is then given by

$$I = \frac{2e}{h} T (\mu_1 - \mu_2). \quad (12.27)$$

When we measure the voltage in a two-wire configuration, i.e., directly at the terminals of the reservoirs injecting the current,  $eV = \mu_1 - \mu_2$  and the conductance is given by

$$G = \frac{I}{V} = \frac{2e^2}{h} T. \quad (12.28)$$

A first observation concerning this relation is that, if we take a sample with perfect transmission, the conductance is not infinite, or the resistance is not zero. There is always a contact resistance  $h/2e^2$  in this configuration.

If the measurement is now made by the four-wire technique, which means the drop in voltage is measured directly at the level of the perfectly conducting wires, the charge balance between the left-hand and right-hand wires gives  $eV = R(\mu_A - \mu_B)$ . We then obtain the Landauer four-wire formula:

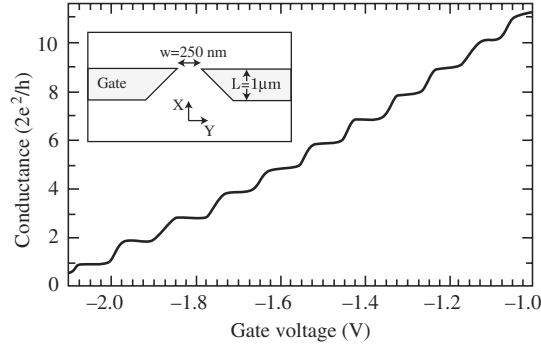
$$G = \frac{I}{V} = \frac{2e^2}{h} \frac{T}{R}. \quad (12.29)$$

The perfect sample ( $T = 1$  and  $R = 0$ ) now has zero resistance.

We see here that the value of the conductance depends on the geometry used in the experiment. The Landauer formula is generalised to the multi-channel case in the following way. We consider the scattering matrix, which is a  $2N \times 2N$  matrix, where  $N$  is the number of channels. Its entries are the probability amplitudes  $T_{ij}$  for a left-hand channel  $i$  to be transmitted into the right-hand channel  $j$ . The generalisation of the Landauer formula is then

$$G = \frac{I}{V} = \frac{2e^2}{h} \sum_i T_i, \quad (12.30)$$

where  $T_i$  are the total transmission coefficients for the various channels, i.e.,  $T_i = \sum_j T_{ij}$ , along the different possible paths from a channel  $i$  to a channel  $j$ .



**Fig. 12.7.** Conductance of an electron gas as a function of the gate voltage which modulates the opening between the two contacts. Steps are observed which correspond to the gradual opening of further channels

For a perfect sample, i.e., one in which the transmission is equal to unity in all channels, it turns out that the conductance is simply given by  $G = Ne^2/h$ . This quantisation of the conductance as a function of the number of channels has been observed experimentally in quantum contact points (see Fig. 12.7). Two gases of electrons with very high mobility are separated by a very narrow opening whose width can be controlled by an electrostatic voltage.

### Weak Localisation

The first quantum interference effect observed experimentally was called weak localisation. The Landauer formula shows that the conductance is related to the transmission  $T_i$  of the various conduction channels. One must sum over all possible paths  $A_\alpha$  to know the total probability of transmission in each of these channels:

$$T_i = \left| \sum_{\alpha} A_{\alpha} \right|^2 = \sum_{\alpha} |A_{\alpha}|^2 + \sum_{\alpha \neq \beta} A_{\alpha} A_{\beta}^*, \quad (12.31)$$

where  $A_{\alpha}$  is the probability amplitude of transmitting channel  $i$  by the path  $\alpha$ . The first term in this sum is the classical term giving the Drude conductance. The second is the interference term. One might think that this cross term is the sum of very different terms which therefore oscillate rapidly and give a negligible contribution to the mean conductance. We shall see that this is not so.

The weak localisation correction concerns rather specific trajectories in the form of loops, i.e., for which the points of departure and arrival coincide. Such a trajectory can be travelled in either direction. If there is time-reversal symmetry, there is no phase difference between these two paths and interference between the two options is constructive. The cross term is the same as the classical term. In other words, the probability of looping the loop is twice as

big if interference effects are taken into account. Since there is a larger probability of going around a loop, the electron propagates less and the resistance therefore increases. All loops with size less than  $L_\varphi$  contribute additively to the correction because the interference effects here are always constructive for each loop.

### 12.3.3 Calculating the Correction

We shall use a semiclassical method due to Khmel'nitskii. We treat classical diffusion as a random walk. The probability of covering a distance  $r$  in time  $t$  is given by

$$P(r, t) = \frac{1}{(2\pi Dt)^{d/2}} \exp\left(-\frac{r^2}{2Dt}\right), \quad (12.32)$$

where  $D$  is the diffusion coefficient. In a space with  $d$  dimensions, a propagating electron occupies a tube of cross-section  $\lambda_F^2$ , where  $\lambda_F$  is the Fermi wavelength. The mean distance travelled during time  $t$  is  $\ell = \sqrt{Dt}$ , and the volume accessible to the electron during this lapse of time is  $V = (Dt)^{d/2}$ . The volume element occupied by an electron is  $\lambda_F^2 v_F dt$ , so the probability of travelling round a loop between times  $t$  and  $t + dt$  is given by the volume ratio

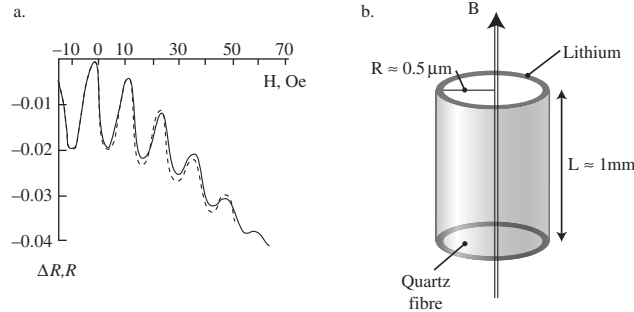
$$P(t)dt = \frac{v_F \lambda_F^2 dt}{(Dt)^{d/2}}. \quad (12.33)$$

The total probability is calculated by integrating this expression between  $t = \tau$ , the elastic collision time, and  $t = t_\varphi$ , the time required to travel a distance  $L_\varphi$ , given by  $L_\varphi = (D\tau_\varphi)^{1/2}$ .

Using the Einstein relation  $\sigma = e^2 \nu D$  and  $\nu = k_F^2 / v_F h$ , we obtain

$$\Delta\sigma = \begin{cases} \frac{2e^2}{h}(L_\varphi - \ell) & \text{in one dimension,} \\ \frac{2e^2}{h} \log \frac{L_\varphi}{\ell} & \text{in two dimensions,} \\ \Delta\sigma = \frac{2e^2}{h} \left( \frac{1}{L_\varphi} - \frac{1}{\ell} \right) & \text{in three dimensions.} \end{cases} \quad (12.34)$$

The weak localisation correction is typically of the same order of magnitude as the conductance of a channel. It will thus become more noticeable as the sample size decreases. Hence, if we measure the sample resistance as a function of temperature, we observe an increase in the resistance at low temperatures which corresponds to the increase in the coherence length and hence to the greater number of loops contributing to interference effects.



**Fig. 12.8.** (a) Experiment of Sharvin and Sharvin. (b) Quartz fibre

### 12.3.4 Effect of Magnetic Fields

The weak localisation correction depends critically on time-reversal symmetry. Now a relatively weak magnetic field will break this symmetry. The magnetic field modifies the the electron action by the introduction of a vector potential. (This amounts to replacing  $\mathbf{p}$  by  $\mathbf{p} + e\mathbf{A}$ .) The phase of the electron wave function is thus modified:

$$\Delta\varphi = \frac{e}{h} \int_C \mathbf{A} \cdot d\mathbf{l} ,$$

where  $C$  is the path travelled. For a closed path,

$$\Delta\varphi = \frac{e}{h} \oint_C \mathbf{A} \cdot d\mathbf{l} = \frac{e}{h} \iint_S \mathbf{B} \cdot d\mathbf{S} = \frac{2\pi\Phi}{\Phi_0} , \quad (12.35)$$

where  $\Phi$  is the flux through the loop and  $\Phi_0 = h/e$  is the flux quantum. We thus obtain a phase change of  $2\pi$  when one flux quantum passes between two trajectories.

In the context of weak localisation, if we only consider a single loop, two trips are required around the loop. We therefore expect the conductance to oscillate as a function of the flux, with a period of  $h/2e$ . This is indeed what was observed by Sharvin and Sharvin in 1988, when they measured the conductance of a cylinder of diameter  $1\mu\text{m}$  composed of quartz coated with lithium (see Fig. 12.8).

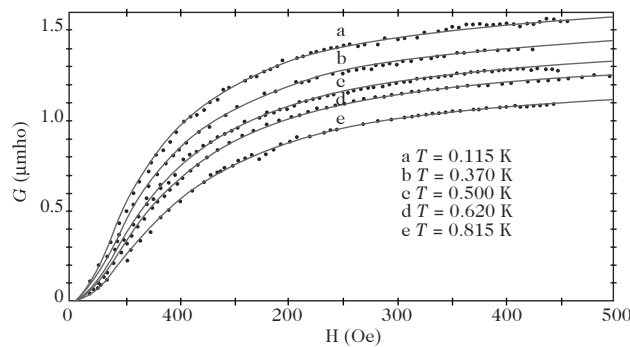
For a sample of arbitrary shape, the magnetic flux modifies the interference between all loops of different sizes and hence mixes up the various periods. The effect of the field is therefore to destroy weak localisation, typically for a value of the field producing a flux of  $\phi_0$  in a loop of size  $L_\varphi$ . Figure 12.9 shows how the conductance of GaAs wires of diameter  $0.7\mu\text{m}$  varies as a function of the magnetic field at different temperatures. Points indicate experimental data and curves correspond to a fitting with theory, where the parameter is  $L_\varphi$ . This is therefore a very good way of measuring the coherence length.

### 12.3.5 Universal Conductance Fluctuations

Consider now a sample with linear dimension smaller than  $L_\varphi$ . The Landauer theory tells us that we must sum over all possible trajectories which transmit a channel  $i$  in order to find the conductance, taking interference terms into account. All pairs of trajectories, like the example shown in Fig. 12.10, will give an interference term whose amplitude depends on the microscopic configuration of the sample, e.g., impurities, grain boundaries, etc.

Surprisingly, the contribution of the interference terms from all pairs of paths does not vanish, but tends to a value which depends on the microscopic configuration. Macroscopically identical samples can thus have a range of different conductances. These fluctuations can be observed in a single sample by using a magnetic field. Indeed, through the vector potential, a magnetic field modifies the phases in a path-dependent way. It is important to note that no mention of time-reversal symmetry is made here. These fluctuations persist even in the presence of very high fields.

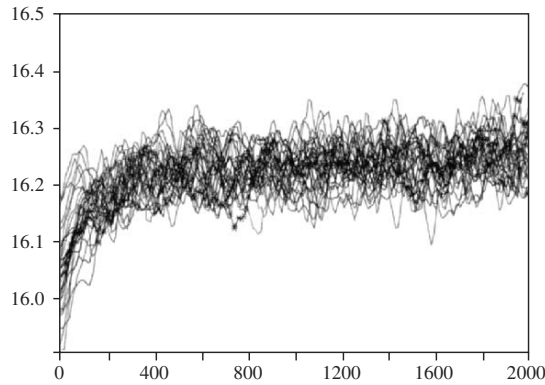
Figure 12.11 shows a set of 50 conductance curves measured on a GaAs wire. Each curve is obtained after applying current pulses to the sample in order to modify the population of ionised impurities responsible for scattering. Following a particular curve as a function of the magnetic field (star), we observe a chaotic variation of the conductance which corresponds to a modification of the interference terms by the magnetic field for a given impurity configuration. Note that each curve is perfectly reproducible and is in fact the



**Fig. 12.9.** Magnetoconductance of GaAs wires at different temperatures, showing the destruction of weak localisation by the magnetic field



**Fig. 12.10.** Example of two trajectories transmitting a channel  $i$



**Fig. 12.11.** Magnetoconductance curves obtained for a GaAs wire after applying electrical pulses to modify the population of ionised impurities

signature of the microscopic configuration of the impurities. This signature is referred to as a magnetic fingerprint.

The average of all these curves would give a signal exactly like the one in Fig. 12.9, i.e., we would retrieve the weak localisation signal. However, if the field is fixed and we follow the variations in the conductance for each curve, we observe the fluctuations due to modifications of the impurities. It is interesting to note that the amplitude of the fluctuations measured for each of these cases – fixed magnetic field or fixed configuration – is the same. This is an instance of the ergodic theorem.

It has been shown that the amplitude of these fluctuations has an astonishing property: it is in fact universal, in the sense that it depends neither on the material, nor on the mean conductance, nor on the number of channels. Indeed,  $\Delta G = e^2/h$ . This is surprising because, if we consider that each conduction channel makes a contribution to the fluctuation, the total fluctuation varies as  $\sqrt{N}$  (fluctuation of  $N$  independent variables). This implies that the channels are correlated. This is in fact a well known result in nuclear physics which follows from the theory of stochastic matrices. It can be summed up rather schematically as follows. The eigenvalues of a matrix whose entries are independent random variables are correlated. In particular, the eigenvalues tend to repel one another. The probability of two eigenvalues being equal is zero. Hence, fluctuations in the conductance are related to fluctuations in the matrix  $T_i$ , which is precisely a stochastic matrix, just as the positions of the impurities are random variables. It is this correlation between the eigenvalues of the transmission matrix which is responsible for the universality of the fluctuations in  $G$ .

### 12.3.6 Cutoffs

What happens when the sample size increases and eventually exceeds  $L_\varphi$ ? We then divide the sample into pieces of linear dimension  $L_\varphi$ . Each piece will give a fluctuation of amplitude  $e^2/h$ , but these fluctuations will not be correlated. The total fluctuation will then be the sum of these independent variables and will vary as  $(L_\varphi/L)^{3/2}$ , because the resistances are summed. It will therefore tend to zero as the sample length increases. We retrieve the classical result.

The temperature is also responsible for a reduction in the amplitude of fluctuations. Consider the thermal length  $L_T = (\hbar D/kT)^{1/2}$ , i.e., the distance that two electrons separated by energy  $kT$  can travel before their phase begins to differ significantly. When  $L$  grows bigger than  $L_T$ , we can apply the same process as above, dividing the sample into  $L/L_T$  independent segments to obtain a variation in the amplitude of fluctuations going as  $T^{-1/2}$ . Increasing temperature also has the effect of reducing the coherence length by favouring collisions with phonons.

## 12.4 An Example of Interference: Aharonov–Bohm Effect

The Aharonov–Bohm effect is an interference phenomenon which arises when an electron trajectory divides into two parts and then joins up again in some region where there is a nonzero vector potential. If phase coherence is preserved, the interference term is proportional to  $\cos \varphi$ , where  $\varphi$  is the phase difference between the two beams, expressed in terms of the vector potential as follows:

$$\begin{aligned}\varphi = \varphi_2 - \varphi_1 &= 2\pi \frac{e}{\hbar} \int_{c_1} \mathbf{A} \cdot d\boldsymbol{\ell} - 2\pi \frac{e}{\hbar} \int_{c_2} \mathbf{A} \cdot d\boldsymbol{\ell} \\ &= 2\pi \frac{e}{\hbar} \oint \mathbf{A} \cdot d\boldsymbol{\ell} \\ &= 2\pi \frac{e}{\hbar} \iint_S \mathbf{B} \cdot d\mathbf{S} = 2\pi \frac{\phi}{\phi_0}.\end{aligned}$$

The intensity of the reconstructed beam thus oscillates as a function of the magnetic flux between the two trajectories. The Aharonov–Bohm effect shows that the vector potential is not a mere mathematical artefact, but a physical reality. Indeed, the electrons need not actually encounter the magnetic flux. It suffices to introduce the flux by means of a very long solenoid placed between the two paths and perpendicular to the plane. Originally observed for electrons in vacuum, the Aharonov–Bohm effect was first demonstrated in a metal in 1985. In this case, due to the size of the system, it is very difficult not to have a magnetic field in the space where the electrons are moving, and an extra signal thus appears.



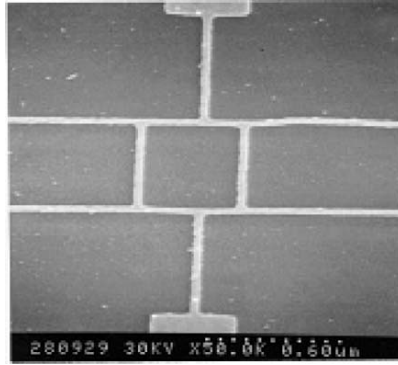


Fig. 12.12. Gold loop

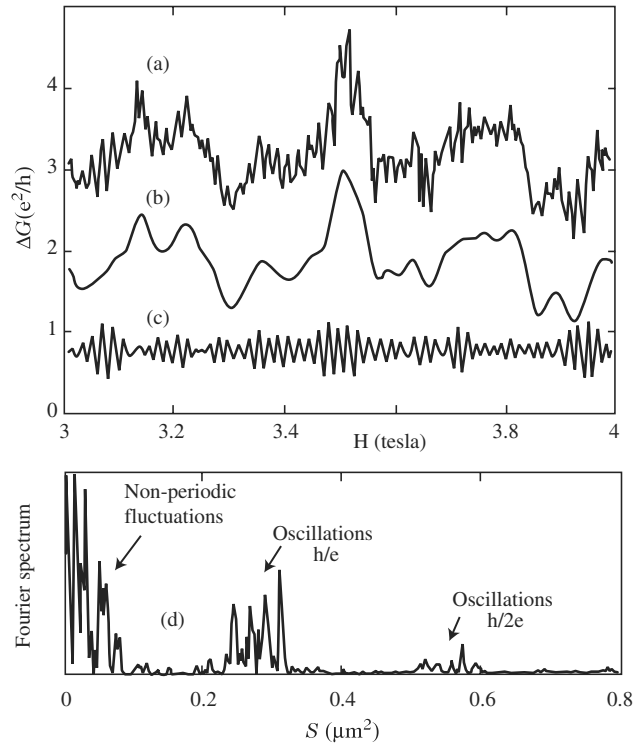
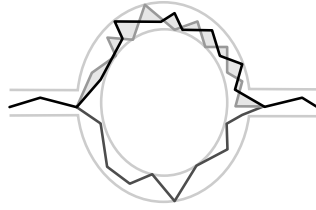


Fig. 12.13. Magnetoconductance of the gold ring and its Fourier transform

Figure 12.12 shows a gold loop of cross-section  $30 \text{ nm} \times 30 \text{ nm}$  and length  $1 \mu\text{m}$ , connected to measuring wires. Such objects can be made by electron lithography (see Chap. 1). If we measure the conductance as a function of the magnetic field for such an object at very low temperatures, so that  $L_\varphi$  is larger than the size of the loop, we observe the signal shown in Fig. 12.13a.



**Fig. 12.14.** The loop can be crossed either by two paths lying in the same branch, or two paths such that one is in each branch

The conductance exhibits periodic oscillations superimposed on slower variations. These two components can be separated by filtering. The fast component shown in Fig. 12.13c is periodic with period corresponding to a flux quantum in the loop. This is the Aharonov–Bohm effect which corresponds to interference between two paths passing through different branches of the loop. The other component, shown in Fig. 12.13b, without a well defined period, corresponds to trajectories which go through the same branch of the loop, which is also penetrated by the magnetic field (see Fig. 12.14). The enclosed area is much smaller and hence corresponds to much bigger variations in the magnetic field. There is no well defined period for this component. It is in fact the expected signal for a wire, revealing the universal fluctuations in the conductance discussed earlier. The Fourier transform of the total signal with the abscissa converted into area does indeed show a signal at  $0.3\mu\text{m}^2$ , which corresponds to the area of the loop. The nonzero width of the peak reflects the nonzero width of the ring. Indeed, the possible periods lie between the inner and outer areas of the ring. A harmonic of the signal with smaller amplitude can also be made out, corresponding to paths which go twice round the ring. Note that this is not the Sharvin–Sharvin oscillation, which would also give a signal of period  $h/2e$ , because the magnetic field is rather strong here (several tesla) and time-reversal symmetry is not established.

### Conclusions

We have attempted to give an introductory discussion of interference effects in nanostructures, together with several relevant examples of observed effects. This is a research theme that has attracted much interest since the 1980s, when the first experiments were carried out. New discoveries are being made on a regular basis. For example, we have not mentioned the effect of interference on thermodynamic properties, an area where quite remarkable results have been found. More recently, noise in nanostructures has begun to emerge as a novel theme in this new branch of physics.

## 12.5 Superconducting Nanoelectronics: RSFQ Logic

### 12.5.1 Introduction

Ever since the 1960s, superconducting nanoelectronics [3,4] has been presented as a possible alternative to semiconductors for making logic components. Indeed, this technology has undeniable advantages for realising ultra-fast components:

- Superconducting transmission lines are non-dispersive, so that ultra-short electrical pulses (of the order of 1 ps) can be transmitted without distortion.
- The switching time of superconducting junctions (Josephson junctions) is of the order of 1 ps.
- The dissipated power in a basic cell is very low.

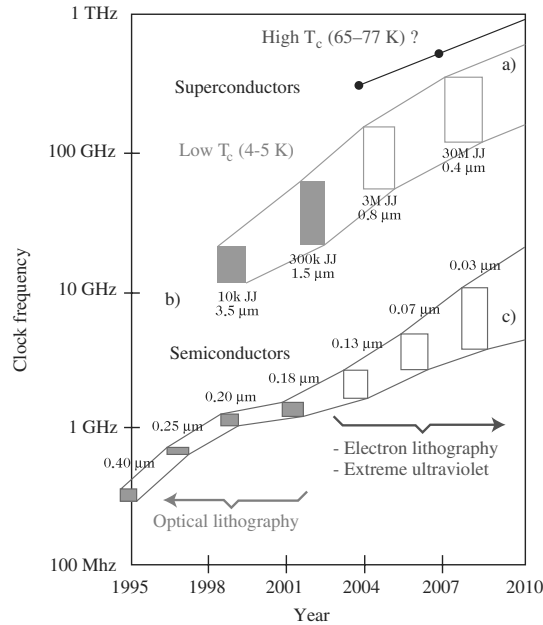
However, despite considerable effort and several large-scale projects, e.g., the IBM superconducting calculator (1969–83) and the MITI project in Japan (1981–90), no practical realisations have been able to compete with traditional approaches to electronics until recently. There are several reasons for this. To begin with, it took a long time to conceive of a superconductor logic able to do anything interesting. Up until the end of the 1980s, operating frequencies were limited to the GHz region, frequencies only just one order of magnitude higher than those in semiconductor circuits. Apart from this, it was difficult to develop a reliable technology for fabricating components, and to make matters worse, during these two decades, semiconductor performance was being improved all the time, so that objectives could always be reached at any given time.

Superconductor electronics based on a new principle – rapid single-flux quantum (RSFQ) logic [6] – has nevertheless come back into the limelight, as a consequence of exceptional performance levels, going well beyond the achievements of mainstream electronics (see Fig. 12.15), which is now approaching its limiting speed. Hence, at the end of the next decade, these circuits may usefully replace circuits based on semiconductors as predicted by the International Technology Roadmap for Semiconductors [5].

### 12.5.2 Superconducting Logic Components

Before the advent of the Josephson junction, core element of all superconducting logic components today, logic devices (e.g., the cryotron) exploited the normal/superconducting phase change to switch between a low resistance (normal) state and a zero resistance (superconducting) state. Such components never led to practical applications due to the intrinsic thermal principles they involved which limited switching speeds.

At the beginning of the 1980s, a family of logic components appeared that were known as latching logic components. These involved switching a



**Fig. 12.15.** Comparison between the semiconductor and superconductor sectors as gauged by the operating frequency of a complex system like a processor. The semiconductor data is taken from the International Technology Roadmap [5] and superconductor data from Table 12.1. (a) High- and (b) low-temperature superconductors. (c) Semiconductors

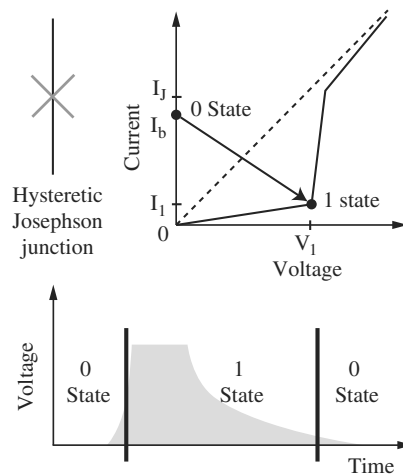
Josephson junction which exhibited a hysteresis effect (see Fig. 12.16). This hysteresis, used to maintain the logic states, nevertheless required the periodic resetting of the bias current of the junction to zero (for a certain time), imposing an intrinsic limit on the clock frequency of the components in the GHz range. This technology never caught on and the technical choice of imitating semiconductor components, encoding logic states by voltages, was brought into question and then rejected.

Rapid single-flux quantum (RSFQ) logic then came onto the scene. It uses shunted, hence non-hysteretic Josephson junctions and exploits their dynamical and quantum behaviour, as the name suggests (see Fig. 12.17). It is not a static voltage level which encodes information in RSFQ components, but rather the presence or absence of a magnetic flux quantum, or fluxon, i.e.,  $\Phi_0 = h/2e = 2.07 \times 10^{-15}$  Wb. The basic RSFQ consists of a superconducting loop closed by a junction and shunted by a resistance. Each change by one flux quantum in the loop (where a fluxon either enters or leaves the loop) induces a voltage pulse across the junction terminals. The time integral of this pulse is  $\Phi_0$ , i.e., 2.07 mV ps. The duration and amplitude of the pulse depend on the geometry of the junction and the material used to make it. For a  $1 \mu\text{m}^2$  niobium (Nb) junction, the pulse lasts for about 1 ps and has amplitude 2 mV.

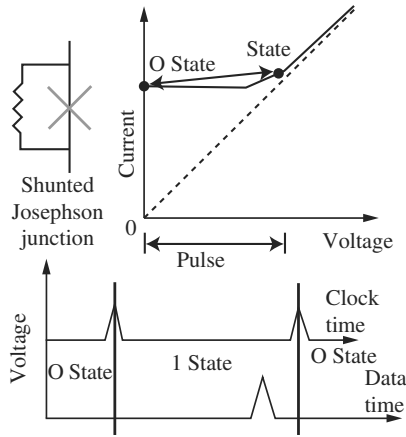
Data processing thus involves manipulating voltage pulses resulting from the transfer of flux quanta, with the great advantage that the energy dissipated during the transfer of a flux quantum is independent of the amplitude of the pulse, being equal to  $I_J\Phi_0$ . For a critical Josephson current of  $100\mu\text{A}$ , this is an energy of  $2 \times 10^{-19}\text{ J}$ , which is five orders of magnitude lower than in a semiconductor.

### 12.5.3 Structure and Performance of RSFQ Components

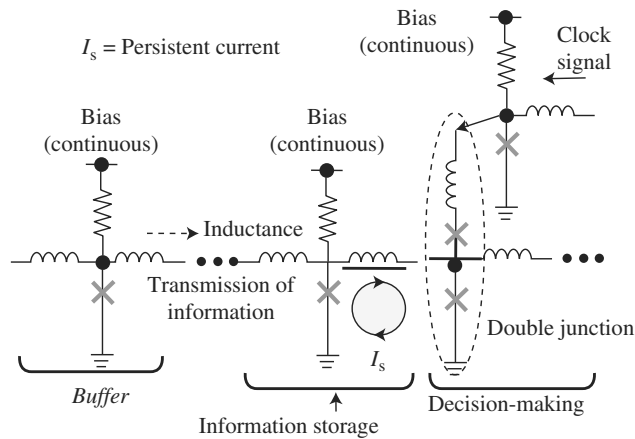
Any RSFQ circuit is made up of three basic cells, as shown in Fig. 12.18 [4]. The first, comprising an inductance and a junction, plays the role of buffer, dealing with the transmission of the picosecond pulses. The second cell, also made up of an inductance and a junction, stores the information (pulse) in the form of a persistent loop current. Finally, the third cell, comprising an inductance and two different junctions controlled by a clock signal, carries out two functions. To begin with, it serves as a buffer cell, preventing a signal arriving at the output from reacting with the input. In addition, it deals with decision-making, in the sense of deciding whether or not to transmit the information (pulse). The inductances associated with the junctions play an integral part in the operation and serve either to couple the various cells together (cell 1), or to store energy (cell 2). Operation is optimised when  $LI_J = \Phi_0/2$ , for  $I_J \sim 100\mu\text{A}$  and  $L \sim 10\text{ pH}$ . For its part, the clock signal serves as a reference for the voltage pulse, to define the logic levels 0 and 1.



**Fig. 12.16.** Underlying principle of latching logic, based on a Josephson junction with hysteretic  $I$ - $V$  characteristic. The bias current  $I_b$  of the junction is fixed at a value slightly below the critical Josephson current  $I_J$ . The junction switches from the superconducting state 0 to a resistive state 1 when the applied bias current momentarily exceeds  $I_J$ . To revert to state 0, the bias current must return to zero for a certain time



**Fig. 12.17.** Underlying principle of RFSQ logic, based on a shunted Josephson junction with a non-hysteretic  $I$ - $V$  characteristic. When the bias current  $I_b$  reaches the critical Josephson value  $I_J$  for the junction, a flux quantum passes through the loop, inducing a voltage pulse at the junction terminals. By convention, the logic state 1 is defined by the presence of a pulse during the period of the clock signal, while state 0 is defined by the absence of such a pulse



**Fig. 12.18.** Basic structure of an RSFQ gate made up of three cells: buffer, information storage, and decision-making

By convention, the logic level 1 is defined by the presence of a pulse during the period of the clock signal, whilst level 0 is defined by the absence of a pulse during this same period.

The performance of RSFQ components is intrinsically higher than the performance of semiconductor components, both for speed and energy dissipation, for a given integration density.

**Table 12.1.** Frequency performance and characteristics of Josephson junctions as a function of their size [7]

$J_J$ [ $\mu\text{A}/\mu\text{m}^2$ ]	$D$ [ $\mu\text{m}$ ]	$C_A$ [FF/ $\mu\text{m}^2$ ]	Operating frequency for a circuit [GHz]		
			Unit	Standard	Complex
10	3.54	50	124	52–78	13–31
20	2.50	54	169	71–106	18–42
40	1.77	58	230	97–145	24–58
100	1.12	64	346	145–217	36–87
200	0.79	70	468	196–294	49–118
500	0.50	80	693	290–435	73–174
1 000	0.35	90	927	388–582	97–330

The uppermost limit for the operating frequency of an RSFQ component carrying out a basic logic function, e.g., a flip-flop, is determined by the width of the pulse created by the fluxon and can be expressed using an RSJ-type (resistively-shunted junction) model [7]:

$$F_J = \frac{\omega_J}{2\pi} = \left( \frac{\beta_c J_J}{2\pi\Phi_0 C_A} \right)^{1/2}, \quad (12.36)$$

where  $\beta_c$  is the McCumber parameter, usually close to unity, determined by the shunt resistance,  $J_J$  is the critical Josephson current density, and  $C_A$  is the specific capacitance of the device with area  $A$ . The thickness of the tunnel barrier is one of the key parameters, like the junction size, because it determines the critical Josephson current density  $J_J$ . As it is reduced,  $C_A$  increases linearly and  $J_J$  increases exponentially, so that the operating frequency  $F_J$  increases as  $\sqrt{J_J}$ . This is why junctions with high values of  $J_J$  are required to make fast components. For asynchronous circuits of medium complexity, operating frequencies lie between  $\omega_J/15$  and  $\omega_J/10$  (empirically), but can be bounded by  $\omega_J/60$  and  $\omega_J/25$  for very complex circuits, such as processors [7]. The frequency ranges given in Table 12.1 and shown graphically in Fig. 12.15 depend only on the design and implantation of the RSFQ components and in no way reflect theoretical physical limits.

The power consumed in an RSFQ logic gate is less than  $\mu\text{W}$ . It is essentially due to the bias current of several hundred  $\mu\text{A}$  which goes through the neighbouring resistances (bias resistance, shunt resistance). In today's components, the contribution due to energy dissipated in the Josephson junction remains negligible (see above, of the order of  $2 \times 10^{-19}$  J). Note that the power cannot be minimised by reducing the bias current, because the product  $\Phi_0 I_J$  must always be much greater than the product  $kT$ , otherwise the noise contribution due to thermal fluctuations is not negligible, and this leads to a prohibitive error/bit ratio for most applications.

**Table 12.2.** Estimated numbers of Josephson junctions for the main digital components. DSP = digital signal processing, SDR = software defined radio

Type of component	Number of junctions [ $10^3$ ]	Applications
Processor, DSP	100–300	Peta-flop computers, SDR
Switching matrix	10–300	Telecommunications
Correlator	10–300	Astronomy, radar
A/D , D/A converter	1–10	Radar, SDR, metrology
Sampler	0.1–1	

The best integration density of (Nb) Josephson junctions achieved so far is  $60 \times 10^3$  JJ/cm<sup>2</sup> [8]. It is not limited by the ultimate dimensions of the junctions, but rather by the shunt resistance which considerably increases the surface area of the basic cell. Indeed, this resistance requires an area of between 50 and 100 times  $D^2$ , where  $D$  is the linear dimension of the junction [9]. Recent work with Nb has shown that junctions measuring less than 300 nm can be made which no longer require the shunt resistance (intrinsic  $\beta_c$  approximately unity) [10]. Moreover, by using planar multilayers, with 8 layers of Nb instead of 3 today, it should be possible to limit the area of an RSFQ gate to  $25D^2$ , giving an integration density comparable with semiconductor components. With a 0.3- $\mu$ m technology, a theoretical density of  $40 \times 10^6$  JJ/cm<sup>2</sup> could be reached, although in practice it will not exceed  $15 \times 10^6$  JJ/cm<sup>2</sup> if thermal aspects are taken into account.

## Conclusion

The most mature technology available today is based on Nb trilayer junctions. However, this material requires operating temperatures in the range 4–5 K and it is intrinsically limited to a response time of 0.7 ps. Note that the fastest circuit to date (frequency of the order of 800 GHz) is a flip-flop device made with 0.25- $\mu$ m niobium technology [11]. Moreover, the most junctions implemented so far was  $90 \times 10^3$  JJ using 1.75- $\mu$ m technology, to make a microprocessor to operate at 20 GHz [8]. An alternative would be to use NbN technology with the advantage of operating at 50% higher speed and at a temperature of the order of 10 K. However, at the present time, it could not be used for applications with more than  $2 \times 10^3$  JJ. Finally, YBaCuO technology, which is still a long way from being operational, has the advantage that non-hysteretic junctions can be made with response time 0.1 ps and operating at temperatures as high as 40–50 K. On the other hand, these high operating temperatures generate extra thermal noise, which must be compensated by a bias current that is ten times as big (since  $T_{YBCO} = 10T_{Nb}$ ). This in turn means using inductances that are ten times as small, on the frontier of our technological capabilities at the present time.



Many electronic functions have already been achieved, e.g., phase-locked loops, D/A and A/D converters, memory cells, auto-correlators, filters, processors, etc. The numbers of junctions required for a given application are evaluated in Table 12.2 [12], but a great deal of technological progress (and financial investment!) remains to be made before RSFQ electronics can go into mass production.

## References

1. Fukuyama, H., and Ando, T. (Eds.): *Transport Phenomena in Mesoscopic Systems*, Springer-Verlag, Berlin (1992)
2. Likharev, K.: Single-electron devices and their applications, Proc. IEEE **87**, 606–632 (1999)
3. Likharev, K.: Superconductors speed up computation, Physics World **10** (5), 39–43 (1997)
4. Brock, D., Track, E., Rowell, J.: Superconductor ICs: The 100 GHz second generation, IEEE Spectrum, 40–46 (Dec 2000)
5. <http://public.itrs.net/>
6. Likharev, K., Semenov, K.: RSFQ logic/memory family: A new Josephson-junction technology for sub-terahertz clock frequency digital systems, IEEE Trans. Appl. Supercond. **1** (1), 3–28 (1991)
7. Kleinsasser, A.: High performance Nb Josephson devices for petaflop computing, IEEE Trans. Appl. Supercond. **11** (1), 1043–1049 (2001)
8. Dorojevets, M., Bunyk, P., Zinoviev, D.: FLUX chip: Design of a 20 GHz 16 bit ultrapipelined RSFQ processor prototype based on 1.75  $\mu\text{m}$  LTS technology, IEEE Trans. Appl. Supercond. **11** (1), 326–332 (2001)
9. Naveh, Y., Averin, D., Likharev, K.: Physics of high Jc Nb/AlO/Nb Josephson junctions and prospects for their applications, IEEE Trans. Appl. Supercond. **11** (1), 1056–1060 (2001)
10. Kadim, A., Mancini, C., Feldman, M., Brock, D.: Can RSFQ logic circuits be scaled to deep submicron junctions?, IEEE Trans. Appl. Supercond. **11** (1), 1050–1055 (2001)
11. Chen, W., Rylyakov, A., Patel, V., Lukens, J., Likharev, K.: Rapid single flux quantum T-flip flop operating up to 770 GHz, IEEE Trans. Appl. Supercond. **9**, 3212–3215 (1999)
12. Tahara, S., Yorozu, S., Kameda, Y., Hashimoto, Y., Numata, H., Satoh, T., Hattori, W., Hidaka, M.: Superconducting digital electronics, IEEE Trans. Appl. Supercond. **11** (1), 463–468 (2001)

## Molecular Electronics

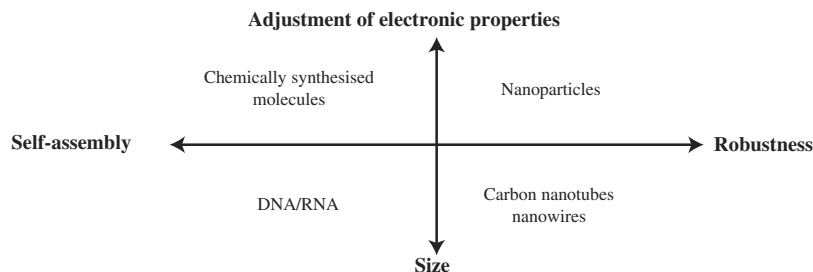
J.P. Bourgoin, D. Vuillaume, M. Goffman, and A. Filoramo

The term ‘molecular electronics’ could be understood in two ways. Firstly, it might be taken as referring to electronics carried out with molecular materials, i.e., organic materials, such as the fabrication of electroluminescent diodes. On the other hand, it could be taken to mean the realisation of devices comprising one or several molecules, or extending the idea slightly, one or several objects of size comparable to a small molecule, i.e., typically about 5 nm. These nano-objects may be organic molecules, metal or semiconductor nanoparticles, carbon nanotubes, nanowires, or biomolecules.

In the present chapter, we shall be concerned primarily with the second interpretation of the term ‘molecular electronics’, which involves synthesising nano-objects endowed with specific functions and connecting them to external electrodes. The aim of molecular electronics is to design and realise electronic circuits using nano-objects as components.

From a fundamental standpoint, the problem is to reach a new transport regime by connecting nano-objects directly to conducting electrodes. By doing this, one may hope to understand and control the relationship between molecular structure and transport properties, in such a way that it will one day be possible to exploit it.

From an applications standpoint, molecular electronics is an idea that has stimulated the efforts and the dreams of a good many scientists, not to mention the commitment of several major industrial partners. But to what do we owe the present enthusiasm for this kind of research? Could it be its profile as a possible alternative to the all-silicon approach to information processing, whose ultimate limits have been announced for a 10 to 15 year horizon? Could it be by reference to the ever-present model of the way our own brain functions? Or is it just an opportunity to shake up an over-conservative microelectronics industry and introduce low cost fabrication methods based on self-assembly of components? Or again, the possibility of developing better control over complexity and reducing the energy cost of computation? Whatever the driving force may be, this recent area of scientific endeavour, by its very nature multidisciplinary, has firmly established itself at the crossroads



**Fig. 13.1.** Basic building blocks of molecular electronics and their position with regard to certain key characteristics

between chemistry, which makes the nano-objects, physics, which studies their properties, and engineering, which fits them together into a working device.

This variety of attributes implies a certain level of complexity that we shall attempt to decode here, starting with the basic building blocks (molecules, but also nanoparticles and nanowires), then describing how they can be connected up to make components, and finally, presenting the methods used to assemble these components into circuits and discussing various possible architectures available at this stage.

### 13.1 Basic Building Blocks: Choice, Wealth, Complexity

The nano-objects providing the basic building blocks of molecular electronics are of four types. Their specific characteristics will be outlined here, insofar as they are relevant to the field of molecular electronics as we have defined it above. For further detail, the reader is referred to other chapters in the present book.

#### Chemically Synthesised Molecules

These molecules, made by chemists, can be adapted to some use for which they have been specifically designed. They can be exactly reproduced and in large quantities, they exhibit quantised electronic levels (allowing observation of quantum effects), they can be bistable or multistable (essential for making components, and especially memory cells), and they are particularly apt for self-assembly. On the other hand, they are very sensitive to their surroundings and generally exhibit a poor thermal behaviour and limited electronic conduction.

#### Biomolecules

Extracted from a biological medium, these molecules are generally rather large (1–100 nm). They often exhibit a strong tendency for self-assembly, and in

particular, can be used as a template for directing the assembly of other objects. Their electronic properties, and in particular those of DNA strands, are still a subject of some controversy today.

### **Metal or Semiconductor Nanoparticles**

Generally protected by an organic matrix, these tiny grains of matter nevertheless conserve their special electronic properties which result from the quantisation of energy levels due to spatial confinement (core size 1–10 nm). More robust and less sensitive to the environment than molecules, their ability to self-assemble is relatively limited and wholly determined by the nature of the organic matrix that contains them. Single-nanoparticle devices such as single-electron transistors have been made recently.

### **Carbon Nanotubes and Nanowires**

These objects, with nanometric diameters and micrometric lengths, may have semiconducting or metallic properties, depending on their structure. They are robust and can withstand heat treatment. On the down side, their purity leaves something to be desired and they have limited ability with regard to self-assembly, essentially determined by the molecules that can be adsorbed onto their outer surface.

In the rest of the chapter, we shall focus more specifically on chemically synthesised molecules and carbon nanotubes.

## **13.2 A Little History**

The use of organic molecules for electronic applications was first suggested in the 1970s. In 1974, A. Aviram and M.A. Ratner [1] published a theoretical study extolling the virtues of electronic components based on molecules comprising electron donor and acceptor systems separated by a saturated organic bridge. It was claimed that such molecules would give rise in an intrinsic manner to a diode-type behaviour, i.e., that electrons would only transfer from the acceptor group towards the donor group (forward direction, as opposed to reverse direction). These molecules, called molecular diodes or molecular rectifiers, would thus exhibit asymmetrical current–voltage characteristics if the donor and acceptor groups could be connected to an anode and a cathode, respectively, to carry out electrical measurements.

Towards the end of the 1970s, F. Carter then suggested other examples of single-molecule components, to be used as triodes, for instance [2].

The first experimental results concerning the molecular diode idea did not materialise for another twenty years after Aviram and Ratner's original suggestion, due to the extraordinary technical difficulties involved [3,4]. During this lapse of time, a whole new field of research had developed around the

theme of chemical synthesis and the characterisation of novel molecular materials with properties that could be applied in the world of industry. Molecules of type donor–ligand–acceptor, and especially mixed valence compounds, were the subject of intense study. These are organometallic complexes containing two metallic centers with two different oxidation levels which constitute the donor and acceptor sites of the molecule. These compounds, usually based upon ruthenium II and III for reasons of stability, are the simplest available molecules for observing intramolecular electron transfer in solution [5]. On the experimental front, transmission [6] and switching [7] phenomena obtained by chemical [8] or photochemical [9] stimuli were observed, and other functions were achieved, such as memory [10], etc. These isolated organic systems, capable of fulfilling basic functions, formed the starting point for what became known as molecular electronics.

However, a major step still had to be taken to make an electrical contact with a single molecule. It was not until the beginning of the 1990s that a solution turned up as a direct consequence of the invention in 1981 of the scanning tunneling microscope by H. Rohrer and G. Binnig (see Chap.3). Indeed, this invention made it possible to establish an electrical contact with a molecule on a solid substrate, opening the way to experimental progress in molecular electronics, with the first measurements on individual molecules in 1995 [11]. Later in the chapter, we shall return to the experimental aspects of the field and their subsequent development, made possible by the introduction of lithographic techniques with resolution suited to molecular dimensions.

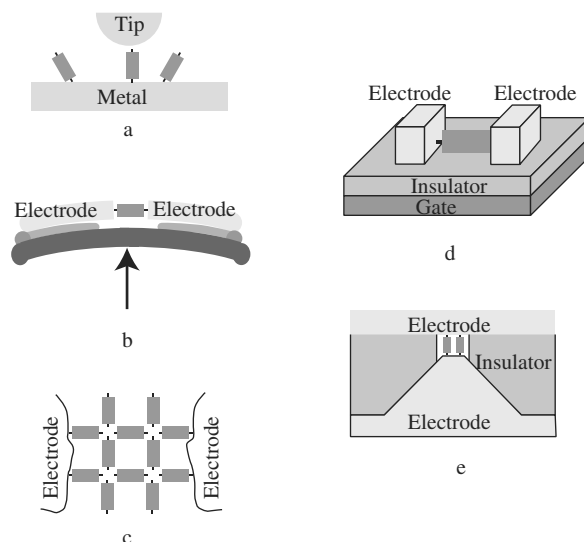
Finally, at the beginning of the 1990s, carbon nanotubes were discovered by Iijima at NEC (Japan), a discovery whose impact on the field of molecular electronics can be clearly measured today.

### 13.3 Molecular Components

This section is concerned with molecular devices, i.e., devices based on the use of one or several electrically-contacted molecules. As mentioned above, the experimental difficulty in contacting one or a small number of molecules is formidable, largely because of the scale reduction by a factor of  $10^7$  to  $10^8$  between the size of the molecule (of nanometric order) and the macroscopic world.

#### 13.3.1 Electrodes and Contacts

Much progress has been made over the past few years in the investigation of electron transport properties in molecules connected to two metal electrodes. This is due to the development of various techniques able to make electrical contacts with one or more molecules. Figure 13.2 shows five experimental arrangements in this context.



**Fig. 13.2.** Experimental arrangements for observation on the molecular scale. (a) The tip of the scanning tunneling microscope is positioned just above the molecule. (b) The molecule is inserted in the gap of a break junction. (c) Molecules serve as bridges between metal nanoparticles in a lattice arrangement. (d) The molecule is in a transistor configuration in a planar nanogap formed on an insulator, with the transistor gate located under the insulator. (e) Molecules are contacted by two metal electrodes in a nanopore arrangement

### Scanning Tunneling Microscopy

Scanning tunneling microscopy (STM) lies at the heart of this development. Indeed, it solves the problem of connecting a molecule by positioning a metal tip above the relevant molecule (see Fig. 13.2a). This technique has made it possible to study the electronic properties of a variety of molecules adsorbed onto various conducting substrates (the reader is referred to reviews in [12, 13]). Four types of experiment can be carried out with this setup:

- The STM tip is positioned just above a molecular monolayer. The number of molecules involved in the measurement is then proportional to the radius of curvature of the tip.
- The molecule under investigation, a conjugated molecule, is inserted into a matrix of molecules that are less transparent to tunnel electrons, e.g., alkylthiols. The result is that the tip, located just above the relevant molecule, seems to ‘stick out’ of the monolayer. Small numbers, or even individual molecules have been successfully observed in this manner [14–17].
- The STM tip is positioned above individual molecules adsorbed onto a substrate coated with an insulating layer. One can then make full use of

the imaging capabilities of the STM and the molecule under measurement can be exactly identified [13].

- A specifically created molecule is adsorbed onto the top of a (double) atomic step of a metal substrate and the STM tip is used to measure the transport along the axis of the molecule. This provides another way of exactly identifying the molecule under measurement [18].

The main conclusion from all these experiments and the theoretical interpretations made of them is that a molecule can have nonzero conductance, determined for the main part by its molecular orbital structure, a point to which we shall return later [19,20].

The use of an STM to connect up a single molecule does have its limitations, however. Among the various problems are the asymmetry of the junction, a lack of mechanical stability which makes it difficult, at least at room temperature, to maintain a stable chemical bond between the tip and the molecule, and the impossibility of working with long molecules (with sizes greater than one nanometer). Moreover, these setups do not permit the introduction of a third electrode which could be used to apply a gate potential. These limitations, combined with technical progress in electron lithography, have led to the development of complementary techniques.

### Complementary Solutions

We shall now describe four of these techniques, all developed over the last few years.

- The first used a mechanical break junction (see Fig. 13.2b). The idea is to break, by bending, a very thin metal wire fabricated on the surface of an elastic substrate, thereby creating two electrodes. Molecules with tethering functions at each end are then introduced between these two electrodes. The interelectrode distance, or gap, is then adjusted to a distance comparable with the size of the molecules in order to establish the contact [15, 21–24]. This technique greatly improves mechanical stability in comparison with STM experiments [23, 25]. It also makes it possible to measure room temperature transport properties of a very small number of molecules chemically bonded to the two electrodes. It is impossible to determine the number of connected molecules precisely, but it can be evaluated as being at most 300 molecules by geometrical considerations and it is reasonable to suppose on the basis of experimental results that only a single molecule comes into play. It is still difficult to build in a third electrode (gate potential) for control purposes.
- A second method consists in using metal nanoparticles connected together by the molecules under investigation (see Fig. 13.2c). The whole thing is then deposited between two metal electrodes whose size, and in particular the radius of curvature, must be compatible with the size of a few

nanoparticles. Measurement of transport properties only yields values averaged over a large number of molecules and it remains difficult to extract the conductance of an individual molecule because the usual laws concerning the conductance are no longer valid on the nanoscale. One thus obtains only an order of magnitude for the conductance of a single molecule [26–30].

- The two last alternatives involve producing a host structure for one or more molecules. Two geometries are possible here:
  - A vertical geometry (see Fig. 13.2e) can be obtained by making a nanopore with diameter 30–50 nm in a thin insulating film deposited on the surface of a metal substrate which constitutes the lower electrode, by self-assembling molecules at the bottom of this nanopore and subsequently evaporating the upper electrode [31–34]. This technique does not allow the electrical connection of a single molecule. The number of molecules present in the nanopore is estimated to be several thousand. Neither does it allow the introduction of a control electrode which could influence transport electrostatically. On the other hand, it is easy to realise and produces interesting results concerning, among other things, the link between transport properties and molecular structure.
  - A horizontal structure (see Fig. 13.2d) can also be obtained, using electron lithography to prepare two metal electrodes, separated by a distance equal to the size of the relevant molecule, on a substrate surface. This setup seems to be the most promising found so far. Indeed, measurements can be made at any temperature and with a control electrode, with possibilities for industrial applications. However, such planar host structures were nevertheless long restricted to the study of molecules with dimensions greater than 5 nm owing to the technical limitations of nanolithography, and in particular the size of the electron beam, the grain size of the electroresistive resist and the grain size of the metal. Over the last few years, several techniques have been devised to go below the 5-nm limit.

The first of these techniques rests upon the following principle. To begin with, two electrodes 20 nm apart are prepared in an SiN film by electron lithography and reactive etching. The two electrodes are then suspended by means of a hydrofluoric acid etch of the lower insulating layer. Finally, the interelectrode spacing is reduced by successive rounds of platinum sputtering, until a gap of 4 nm is obtained [35]. Although this technique does not allow a control electrode to be placed in the immediate vicinity of the relevant molecule, it has nevertheless made it possible to measure the transport properties of palladium clusters [35] and a DNA strand [36], for example.

The second technique was developed by C. Marcus and coworkers at Stanford [37, 38]. The starting point is once again to fabricate two electrodes, but this time 50–400 nm apart. The interelectrode spacing is then reduced down to about 1 nm by gradual electrodeposition of a metal onto the two



electrodes. One advantage with this process is to be able to obtain two electrodes made from different metals, by carrying out two successive rounds of deposition with different electrolytes. Moreover, the metal deposits can be controlled by adjusting the current passing between the two electrodes. As a consequence, the distance between the two electrodes can be controlled to an accuracy of atomic order. This technique has not yet been able to measure the transport properties of conjugated molecules, but it has been used to reveal dynamical Coulomb blockade effects with an alkyldithiol layer inserted in the junction [38].

The third process was devised by McEuen and coworkers [39,40]. The idea is to use a bilayer of electrosensitive resist to make a PMMA bridge suspended over the substrate, then to evaporate gold from two opposing angles calculated in such a way that the two half wires created under the mask overlap to yield a continuous gold wire. The next step is to generate an electromigration process, wherein the gold atoms making up the wire are set in motion by applying a bias across the terminals, until the wire eventually ruptures. This forms two electrodes separated by 1–3 nm. The technique has been used recently to measure the transport properties of  $C_{60}$  molecules [40], among other things.

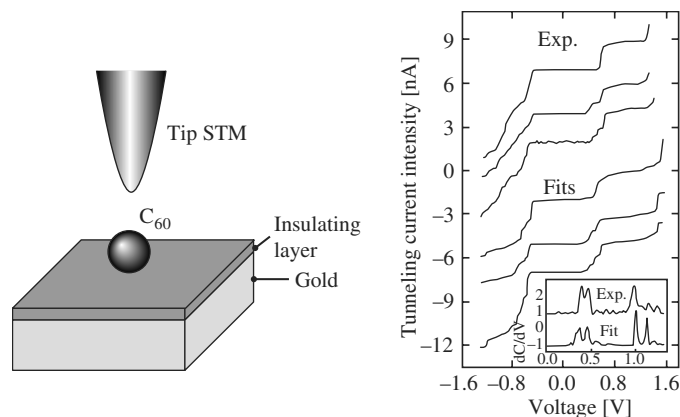
This last process has no major drawbacks, except for the possibility of creating metal clusters at the instant when the wire ruptures, the high temperature of the wire at this same moment [41], and the uncertainty regarding the number of molecules adsorbed within the gap. Indeed, the interelectrode separation cannot be observed by high resolution transmission electron microscopy (HRTEM) and the organic molecules are not visible with such a technique.

### Molecule–Metal Coupling

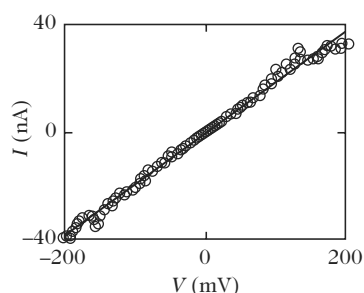
Molecule–metal coupling plays a predominant role in transport within metal–molecule–metal junctions. The example of  $C_{60}$  will serve to illustrate.

$C_{60}$  molecules, discovered in 1985, are spherical molecules of diameter 7 Å. They have been the subject of many studies, in particular by STM after adsorption onto a substrate. It transpires that the transport mechanism can vary greatly with the strength of the electron coupling between molecule and substrate.

Porath et al. used a gold substrate coated with a thin insulating layer [42,43]. The  $I(V)$  characteristic and tunnel spectroscopy results were obtained for an isolated  $C_{60}$  molecule at room temperature and at 4.2 K. It was observed that the current was suppressed at low values of the potential, indicating the presence of the Coulomb blockade phenomenon. Steps were also observed in the current (Coulomb stairs). Figure 13.3 shows examples of these features together with the corresponding theoretical curves, calculated using a classical single-electron transport model modified to take into account the discrete energy spectrum of the molecule. The interpretation of the experimental results



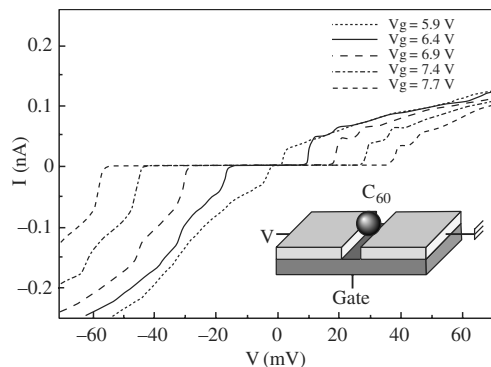
**Fig. 13.3.**  $I(V)$  transport measurements made by STM on  $C_{60}$  in a weak coupling configuration. Courtesy of Porath et al. [42]. The experimental  $I(V)$  characteristics are shown in the *three upper curves*, while the *three lower curves* are simulations. *Insert:* Derivative of the current with respect to the voltage, showing good agreement between experiment and simulation



**Fig. 13.4.** Example of an  $I(V)$  curve recorded by an STM tip placed above a  $C_{60}$  molecule adsorbed onto a gold surface [19]

leads to the conclusion that the  $C_{60}$  molecule is in a double tunnel junction configuration, i.e., with weak metal–molecule coupling.

The situation is very different if the  $C_{60}$  molecule is deposited directly on the metal substrate and there is no insulating layer. Joachim et al. [19, 44] have investigated this configuration: an STM tip is positioned just above a single  $C_{60}$  molecule adsorbed onto a gold (110) surface. Figure 13.4 shows that the  $I(V)$  characteristics are linear for low biases. The resistance of the tip– $C_{60}$ –Au(110) junction is then equal to  $54.8\text{ M}\Omega$ . The authors concluded from these observations that the direct adsorption of the  $C_{60}$  molecule onto the surface gives rise to a situation in which the coupling is strong enough for the tunneling current to cross the junction coherently, but weak enough for the molecule to be able to maintain its electronic identity. The linearity of  $I(V)$  thus results from a broadening of the molecular energy levels due to



**Fig. 13.5.** Examples of  $I(V)$  curves recorded across the terminals of a  $C_{60}$  molecule inserted into a junction prepared by electromigration [40]

interaction with the metal surface (see Sect. 13.3.2) and also from the absence of any resonance between the various levels in the considered energy range. The authors also showed that an electromechanical component could be made on the basis of this transport mechanism.

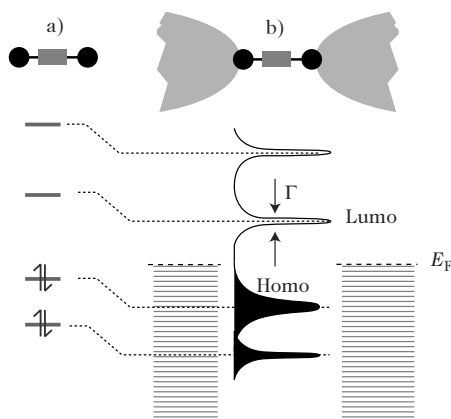
It would be a mistake to conclude that  $C_{60}$  molecules can simply be adsorbed onto a gold surface in order to obtain a strong enough coupling to give rise to a tunnel transport regime in the junction. A recent experiment by McEuen and coworkers [39, 40] shows that this is not the case. This work was the first experimental realisation of a single-molecule transistor. It comprised a nanoscale gap, prepared by controlled electromigration (see p. 454), into which a  $C_{60}$  molecule was introduced. The  $I(V)$  characteristics reveal non-linear behaviour with suppressed current at weak bias (see Fig. 13.5). These results seem to indicate that single-electron transport takes place in the junction, as would characterise a weak coupling between the  $C_{60}$  molecule and the electrodes.

These examples show that the coupling between  $C_{60}$  molecules and a substrate is a delicate problem that we are far from being able to control. A priori, a better way to control this coupling is to use molecules with functions at their extremities that have been specifically chosen to react with the electrode metal, and indeed, this line of research has been widely investigated.

### 13.3.2 Relationship Between Molecular Structure and Properties

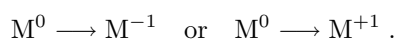
The relationship between molecular structure and transport properties clearly depends heavily on the type of coupling between the molecule and the electrode metal. There are two limiting cases:

- The strong coupling limit, in which the metal–molecule–metal structure forms a sort of electronic wave guide (coherent transport).



**Fig. 13.6.** Energy level diagram. (a) Molecule. (b) Metal–molecule–metal structure for a molecule strongly bound to the electrodes. Peaks derived from the molecular orbitals correspond to the local density of states

- The weak coupling limit, in which the passage of an electron from one electrode to the other occurs sequentially, passing through a charged state of the molecule, viz.,



In the following, we shall present a simple model for each limiting case in order to obtain a qualitative understanding of molecular conduction.

To understand the current–voltage ( $I$ – $V$ ) characteristic of a molecular conductor, the first step is to construct the energy level diagram which allows one to distinguish the weak and strong coupling cases.

The electronic properties of an isolated molecule can be described in terms of the molecular orbitals (MO) [45]. In general, the energy spectrum can be obtained by two different approaches: an ab initio method or a semi-empirical method. In both cases, we obtain a spectrum like the one shown schematically in Fig. 13.6a. The highest occupied molecular orbital is referred to as the HOMO and the lowest unoccupied molecular orbital as the LUMO. This spectrum is modified when the molecule is directly chemisorbed onto the metal electrodes. We shall assume here that the molecule retains its identity during adsorption, in the sense that it does not dissociate, for example. The energy levels of its MO are broadened due to hybridisation with the delocalised wave functions of the metal electrodes (see Fig. 13.6b). These energy levels can also be shifted energywise by the interaction with the electrodes under the effects of fractional charge transfer.

It is the degree of broadening of the levels which will allow us to distinguish between weak and strong coupling. To do this, we compare this broadening  $\Gamma$

with the energy  $U$  required to add or subtract an electron from the molecule adsorbed on the electrodes:<sup>1</sup>

- If  $U \leq \Gamma$ , transport will be described by the strong coupling case.
- If on the other hand  $U \gg \Gamma$ , we find ourselves in the weak coupling limit, where the passage of an electron from electrode 1 to electrode 2 occurs sequentially via a charged state of the molecule ( $M^0 \rightarrow M^{-1}$  or  $M^0 \rightarrow M^{+1}$ ). This regime, known as the Coulomb regime, is characterised by integer charge transfer. It is important to note that, in order for the structure to be in the Coulomb regime, both electrodes must be weakly coupled to the molecule, because  $\Gamma$  is the sum of the individual broadening due to each electrode.

### Transport in the Strong Coupling Regime

Intuitively, one expects the coupling force to be an important factor in determining the current through these structures. Indeed, the stronger the coupling, the greater one expects the current flow to be.

The broadening  $\Gamma$  of a molecular orbital which reflects the coupling force can also be associated with the escape time  $\tau$  of an electron placed in this molecular orbital:  $\Gamma = \hbar/\tau$ . One can also interpret  $\Gamma/\hbar$  as the injection rate into the molecular orbital once contact has been made. In general, the broadening  $\Gamma$  can be different for different molecular orbitals. Two quantities  $\Gamma_1$  and  $\Gamma_2$  are usually defined, one for each contact, with a total broadening given by  $\Gamma = \Gamma_1 + \Gamma_2$ .

To understand the way the current passes through a molecule in the strong coupling regime, two things are required:

- The energy diagram of the molecular orbitals must be adjusted with respect to the Fermi level of the electrodes.
- The spatial profile of the applied potential must be established.

#### *Energy Diagram*

The position of the Fermi level with respect to the HOMO and LUMO of the molecule is probably the single most important factor in determining the  $I(V)$  characteristic of a molecular conductor. In general, it is located between the HOMO and LUMO of the molecule.<sup>2</sup> The position of the Fermi level can be determined by photoemission spectroscopy [46].

<sup>1</sup> The energy term  $U$  for an isolated molecule is easy to determine using standard methods of quantum chemistry. For a molecule adsorbed onto the metal electrodes, it is more difficult to determine and indeed, this question is the subject of current research.

<sup>2</sup> The Fermi energy  $E_F$  is determined by the condition which establishes that the number of states below the Fermi level is equal to the number of electrons of the molecules (multiplied by 2 to account for spin degeneracy). This quantity is not necessarily a whole number when the molecule is in contact with the metal

*Potential Profile*

An important factor in the determination of the  $I(V)$  characteristic is the potential profile across the molecular conductor. At equilibrium, the metal–molecule–metal system has a common Fermi energy, equal to the electrochemical potential  $\mu_1$  and  $\mu_2$  of the two electrodes. When a potential difference  $V$  is applied across the structure,  $\mu_1 - \mu_2 = eV$ . We now ask how the electrochemical potentials  $\mu_1$  and  $\mu_2$  will evolve with respect to the molecular levels in this case.

Of course, we are free to choose an arbitrary zero level for the applied potential. For example, if electrode 1 is taken as reference,

$$\mu_1 = E_F, \quad \mu_2 = E_F + eV.$$

However, we must also account for the shift in the energy of the molecular orbitals, which depends in detail on the shape of the electrostatic potential profile in the molecule. To a first approximation, the molecular levels can be considered to shift in a rigid manner by a change in the average potential  $\langle \delta V_{\text{mol}}(r) \rangle$  in the molecule in the presence of an applied bias  $V$  [47].

In this approximation, we may write the average potential as

$$\langle \delta V_{\text{mol}}(r) \rangle = \eta eV,$$

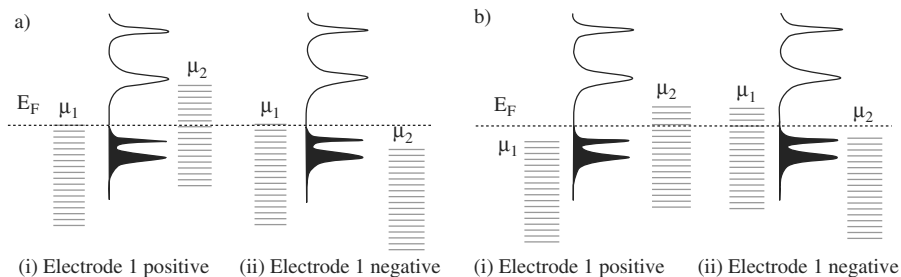
where the factor  $\eta$  multiplying the voltage is a number between 0 and 1. For convenience, let us now take the HOMO energy as reference. The electrochemical potentials are then shifted as follows:

$$\mu_1 = E_F - \eta eV, \quad \mu_2 = E_F + (1 - \eta)eV.$$

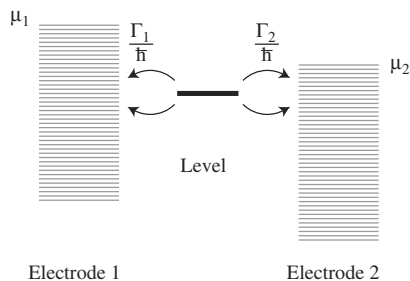
It is important to note that this fractional voltage factor  $\eta$  can have a profound effect on the  $I(V)$  characteristic. If  $\eta = 0$ , the energy diagram under an applied voltage looks like the one shown schematically in Fig. 13.7a. For a positive potential applied at electrode 1, the current begins to circulate when  $\mu_2$  reaches the LUMO level, whereas for a negative potential, it begins to circulate when  $\mu_2$  reaches the HOMO. As a consequence, the positive branch of  $I(V)$  can be completely different from the negative branch, because different molecular orbitals contribute to transport. On the other hand, if  $\eta = 0.5$ , the energy diagram is like the one shown in Fig. 13.7b and  $I(V)$  will be symmetrical. Indeed, independently of the polarity of the applied bias, when the bias is increased, conduction will first take a contribution corresponding to the HOMO (resp. LUMO) and the LUMO (resp. HOMO) will only contribute at high voltages, if the equilibrium Fermi energy  $E_F$  is close to the HOMO (resp. LUMO).

---

electrodes. The molecule may give or take a fractional charge  $\delta n$  depending on the work function of the metal electrons. For most metal–molecule bonds,  $\delta n < 1$ .



**Fig. 13.7.** Energy diagram for a metal–molecule–metal structure. (i) Electrode 1 positive with respect to electrode 2. (ii) Electrode 1 negative with respect to electrode 2. (a)  $\eta = 0$  and molecular levels remain fixed with respect to electrode 1. (b)  $\eta = 0.5$  and molecular levels are shifted by  $eV/2$  with respect to electrode 1



**Fig. 13.8.** Schematic representation of the of the model with one molecular energy level.  $\Gamma_1$  and  $\Gamma_2$  are the couplings of the level with electrodes 1 and 2, respectively

### Simple Model

We shall describe here a simple model with a single molecular orbital of energy  $\varepsilon$ . This level represents the level closest to the Fermi energy  $F_F$  at equilibrium. This single-level model incorporates the main ingredients outlined in the last section and illustrates the roles played by each of these factors:

- the position of the Fermi level  $E_F$  relative to  $\varepsilon$ ,
- the broadening  $\Gamma_1$  and  $\Gamma_2$  due to the electrodes,
- the charge energy  $U$  which produces the energy level shift under applied bias  $V$ .

For the moment, we shall neglect the broadening of the level  $\varepsilon$  and treat it as being discrete. The current through this level can be calculated using a straightforward approach (see Fig.13.8). Let  $\Gamma_1$  and  $\Gamma_2$  be the coupling

---

If  $\delta n = 1$ , the Fermi energy would be located at the LUMO of the molecule, whereas if  $\delta n = -1$ , the Fermi energy would be located at the HOMO. Clearly, for intermediate values, the Fermi level is located between the HOMO and LUMO of the molecule.

between the molecular level and electrodes 1 and 2, respectively. As mentioned earlier, the escape rates of electrons in the level to electrodes 1 and 2 are given by  $\Gamma_1/\hbar$  and  $\Gamma_2/\hbar$ , respectively.

If the level were in equilibrium with electrode 1, the number of electrons  $N_1$  occupying it would be

$$N_1 = 2f(\varepsilon, \mu_1) ,$$

where

$$f(\varepsilon, \mu) = \left[ 1 + \exp\left(\frac{\varepsilon - \mu}{k_B T}\right) \right]^{-1}$$

is the Fermi–Dirac function. In the same way, if the level were in equilibrium with electrode 2, this number would be

$$N_2 = 2f(\varepsilon, \mu_2) .$$

Out of equilibrium, the number of electrons  $N$  will lie between  $N_1$  and  $N_2$  and we can write the net current through electrode 1 as

$$I_1 = \frac{e\Gamma_1}{\hbar}(N_1 - N) ,$$

whilst for electrode 2,

$$I_2 = \frac{e\Gamma_2}{\hbar}(N - N_2) .$$

In the stationary state,  $I_1 = I_2$  and hence

$$N = 2 \frac{\Gamma_1 f(\varepsilon, \mu_1) + \Gamma_2 f(\varepsilon, \mu_2)}{\Gamma_1 + \Gamma_2} ,$$

$$I = I_1 = I_2 = \frac{2e}{\hbar} \frac{\Gamma_1 \Gamma_2}{\Gamma_1 + \Gamma_2} \left[ f(\varepsilon, \mu_1) - f(\varepsilon, \mu_2) \right] .$$

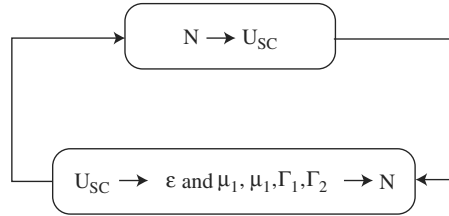
Given the energy  $\varepsilon$  of the molecular level, the couplings  $\Gamma_1$  and  $\Gamma_2$  with the electrodes, and the electrochemical potentials  $\mu_1$  and  $\mu_2$  of the electrodes, we can obtain the current from the last equation. However, we would like to include charge effects in our calculation. To do so we shall introduce a potential similar to the one used in the Hubbard model, viz.,  $U_{\text{SC}} = \langle \delta V_{\text{mol}}(r) \rangle$ , due to the change in the number of electrons from the equilibrium value of  $2f(\varepsilon_0, E_{\text{F}})$ :

$$U_{\text{SC}} = U \left[ N - 2f(\varepsilon_0, E_{\text{F}}) \right] .$$

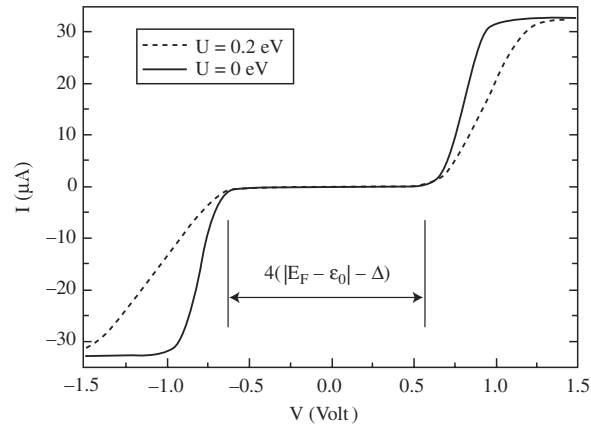
We shall allow the level  $\varepsilon$  to float, so to speak, with this potential:

$$\varepsilon = \varepsilon_0 + U_{\text{SC}} .$$





**Fig. 13.9.** Principle of self-consistent calculation



**Fig. 13.10.** Current–voltage characteristic  $I(V)$  calculated from the model with  $E_F = -5.3$  eV,  $\epsilon_0 = -5.7$  eV,  $\Gamma_1 = 0.1$  eV,  $\Gamma_2 = 0.2$  eV

As the potential depends on the number  $N$  of electrons, the calculation must be carried out self-consistently, as shown schematically in Fig. 13.9.

Once convergence has been obtained, the current is calculated using the corresponding expression. For example, the result obtained for  $E_F = -5.3$  eV,  $\epsilon_0 = -5.7$  eV,  $\Gamma_1 = 0.1$  eV,  $\Gamma_2 = 0.2$  eV is shown in Fig. 13.10 with ( $U = 0.2$  eV) and without ( $U = 0$  eV) charge effects. The width of the observed step (with  $U = 0$  eV) is due to the temperature used in the calculation ( $k_B T = 0.025$  eV), because the level is considered to be discrete in this model. Note that including charge effects broadens the steps, even if the broadening effects due to  $\Gamma_1$  and  $\Gamma_2$  are not included in this calculation. The size of the zero current region is directly related to the energy difference between the molecular level and the Fermi level. The current increases significantly when the voltage reaches 0.8 V, i.e., exactly  $2|E_F - \epsilon_0|$ , as one might expect if  $\eta = 0.5$  (see Fig. 13.7b).

One remarkable effect is the asymmetry observed in  $I(V)$  when  $\Gamma_1 \neq \Gamma_2$  (see the curve corresponding to  $U = 0.2$  eV in Fig. 13.10). This can explain several experimental results which show an  $I(V)$  asymmetry [23,48], discussed in more detail on p. 473. In the example shown in Fig. 13.10, the current is

transported by the HOMO level ( $E_F > \varepsilon_0$ ) and it is greater when a positive bias is applied to the better coupled electrode.

#### *Model Including Level Broadening*

In the last section, we treated the level  $\varepsilon$  as discrete, ignoring the broadening  $\Gamma = \Gamma_1 + \Gamma_2$  caused by coupling to the electrodes. In order to take this into account, we replace the discrete level by a Lorentzian density of states

$$D(E) = \frac{1}{2\pi} \frac{\Gamma}{(E - \varepsilon)^2 + (\Gamma/2)^2}$$

and modify the equations for  $N$  and  $I$  to include an integration over the energy:

$$N = 2 \int_{-\infty}^{\infty} dE D(E) \frac{\Gamma_1 f(E, \mu_1) + \Gamma_2 f(E, \mu_2)}{\Gamma},$$

$$I = \frac{2e}{\hbar} \int_{-\infty}^{\infty} dE D(E) \frac{\Gamma_1 \Gamma_2}{\Gamma} [f(E, \mu_1) - f(E, \mu_2)].$$

The charge effect is introduced as before, allowing the centre  $\varepsilon$  of the density of states to float:

$$\varepsilon = \varepsilon_0 + U_{SC}, \quad U_{SC} = U [N - 2f(\varepsilon_0, E_F)].$$

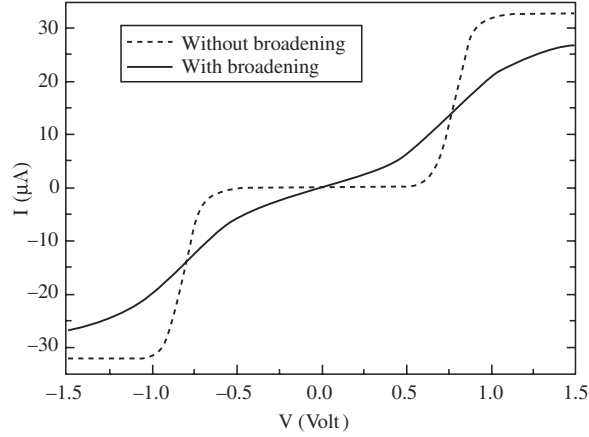
The results are shown in Fig. 13.11, where we have used the same parameters as in Fig. 13.10. The only observed effect is that the steps in  $I(V)$  are now rounded off.

The model just presented includes the three factors which influence molecular conduction, viz.,  $E_F - \varepsilon_0$ ,  $\Gamma_{1,2}$ , and  $U$ . However, real molecules have several molecular orbitals which broaden and may overlap. One of the most widely used formalisms for dealing with several levels manifesting arbitrary broadening and overlap is the non-equilibrium Green function approach. The interested reader is referred to [49].

### **Transport in the Weak Coupling Regime**

#### *Phenomenological Description*

Consider the case of a molecule placed between metal electrodes, to which it is weakly coupled. We shall assume that, in this limit, the number  $N$  of electrons in the molecules is a good quantum number. The coupling is then treated as a perturbation and electron transfer between electrode 1, the molecule and electrode 2 takes place via the tunneling effect.



**Fig. 13.11.** Current–voltage characteristic  $I(V)$  calculated with  $E_F = -5.3$  eV,  $\varepsilon_0 = -5.7$  eV,  $\Gamma_1 = 0.1$  eV,  $\Gamma_2 = 0.2$  eV. *Continuous curve*: Broadening included in the calculation. *Dashed curve*: Broadening not included (identical to the continuous curve in Fig. 13.10)

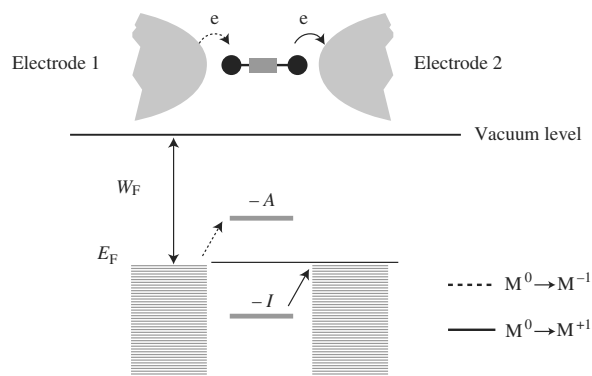
In this framework, the energy diagram is as shown in Fig. 13.12, where we have aligned the energy levels using the work function  $W_F$  of the electrons, the electron affinity  $A$ , and the ionisation potential  $I$  of the molecules. For example, the electron work function for a gold electrode is typically about 5.3 eV, whilst the electron affinity and ionisation potential for an isolated  $C_{60}$  molecule are  $A_0 = 2.65$  eV and  $I_0 = 7.58$  eV, respectively. These values corresponding to emission from and injection into the vacuum are certainly modified by the presence of the metal electrodes. For example, the true values  $A$  and  $I$  differ from  $A_0$  and  $I_0$  by the image potential  $W_{im}$  associated with the metal electrodes [50]:

$$A = A_0 + W_{im} , \quad I = I_0 + W_{im} .$$

The key point in this description of the transport is as follows. When transferred from electrode 1 to electrode 2, the electron must:

- either begin by charging up the molecule ( $M^0 \rightarrow M^{-1}$ , dashed curve in Fig. 13.12) from electrode 1, which requires an energy  $A - W_F$ , and subsequently move to electrode 2, which corresponds to an energy gain;
- or begin by ionising the molecule towards electrode 2 ( $M^0 \rightarrow M^{+1}$ , continuous curve in Fig. 13.12), which involves an energy of  $W_F - I$ , then compensate for this charge transfer from electrode 1, which is energetically favourable.

If the sum of the thermal energy  $k_B T$  and the energy from the voltage supply  $V$  is not enough to overcome  $A - W_F$  or  $W_F - I$ , the current will be blocked, a phenomenon known by the name of Coulomb blockade (see Chap. 11).



**Fig. 13.12.** Energy level diagram of the metal–molecule–metal structure for a weakly coupled molecule

Transport through this metal–molecule–metal device is completely analogous to the problem of a quantum dot connected to electrodes through tunnel barriers (see Chap. 11). The electrochemical potential of the molecule in its neutral state (with  $N$  electrons) is defined as

$$\mu_{\text{mol}}(N) \equiv E(N) - E(N - 1) ,$$

where  $E(N)$  is the total energy of the molecule modified by the presence of the electrodes. This is precisely the definition of the ionisation potential:

$$-I = E(N) - E(N - 1) .$$

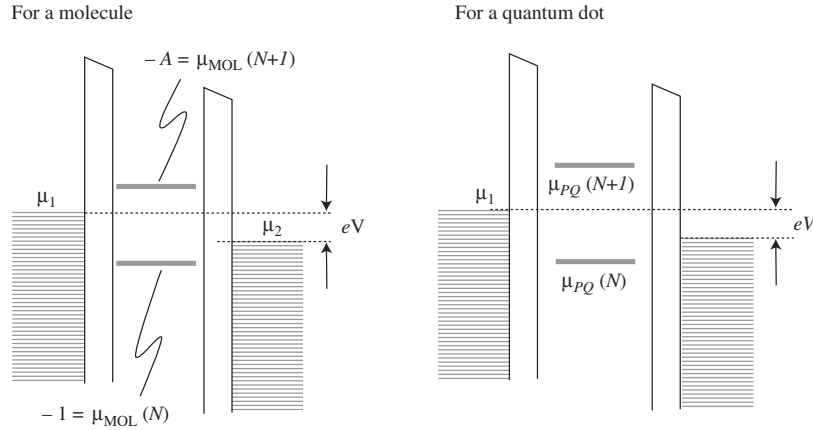
In the same way, the electrochemical potential of the negatively charged molecule (with  $N + 1$  electrons) is equal to

$$\mu_{\text{mol}}(N + 1) \equiv E(N + 1) - E(N) = -A .$$

As for a quantum dot (see Fig. 13.13), the electrons will be able to go through electrode 1 towards electrode 2 when  $\mu_{\text{mol}}$  is located between the potentials  $\mu_1$  and  $\mu_2$  of the electrodes (with  $eV = \mu_1 - \mu_2$ ), i.e.,  $\mu_1 > \mu_{\text{mol}} > \mu_2$ .

#### *Calculating the $I(V)$ Characteristic in the Coulomb Regime*

This analogy can be taken further. The same formalism can in fact be used to calculate the  $I(V)$  characteristic of this metal–molecule–metal structure. To do this, we have represented the equivalent circuit diagram for this structure in Fig. 13.14a. For this equivalent circuit, the total energy of the system can be written as the sum of two contributions: an electrostatic energy term in which the Coulomb interaction between electrons is characterised by the capacitance  $C_i$  between the electrode  $i$  and the molecule, and a second term



**Fig. 13.13.** Analogy between the energy level diagram of a metal–molecule–metal structure for a weakly coupled molecule and a quantum dot (see Chap. 11). The applied voltage  $V$  opens an energy gap  $\mu_1 - \mu_2$  between occupied states in electrode 1 and unoccupied states in electrode 2. The electrons in this range are those contributing to the current. At low voltage only the ground state will contribute to the current [ $\mu_{\text{mol}}(N)$  for the molecule,  $\mu_{\text{QD}}(N)$  for the quantum dot]. At higher voltages, there will also be a contribution from excited states

which characterises the total energy of the molecule. In this approximation, the total energy of the system can be written

$$E(N) = \frac{e^2}{2C}(N - N_0)^2 + \sum_{i=1}^N \varepsilon_i,$$

where  $C = C_1 + C_2$  is the total capacitance between the molecule and the electrodes, which characterises the Coulomb interaction, and  $N_0$  represents a shift charge associated with the environment of the molecule. The last term of the equation is the sum over occupied states of the discrete spectrum  $\varepsilon_i$  of the isolated molecule, calculated using the single-electron approximation.<sup>3</sup>

The electrochemical potential of the molecule with  $N$  electrons is given by

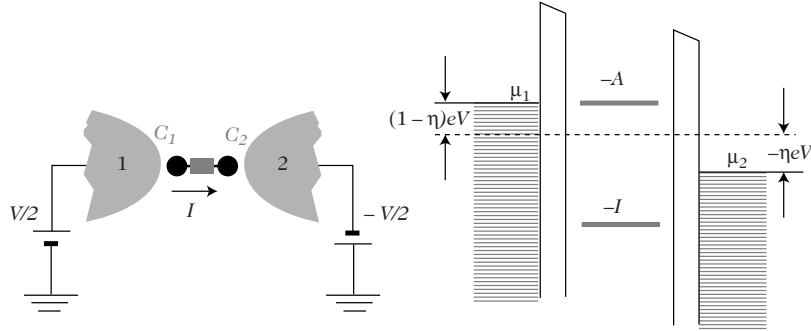
$$\mu_{\text{mol}}(N) \equiv E(N) - E(N-1) = \frac{e^2}{C} \left( N - N_0 - \frac{1}{2} \right) + \varepsilon_N.$$

The electrochemical potential of the electrodes can be written (strong coupling as discussed above)

$$\mu_1 = (1 - \eta)eV, \quad \mu_2 = -\eta eV,$$

where  $\eta = C_1/C$  is the coupling asymmetry factor.

<sup>3</sup> Here we neglect the correlation and exchange term which does not usually significantly modify the energy associated with addition of one electron to the molecule.



**Fig. 13.14.** Equivalent circuit diagram. The molecule is placed between the two electrodes and connected to them via capacitances  $C_1$  and  $C_2$ . Electrons can pass between 1 and the molecule and between 2 and the molecule by the tunnel effect

In the following, the  $I(V)$  characteristic is calculated in the very simple case of a molecule which has only two quantum states: the neutral state (with  $N$  electrons) and a state in which it is charged by an extra electron (hence, with  $N + 1$  electrons).

The transition rate with which electrons pass through the two metal-molecule tunnel barriers in the two directions are calculated to first order in perturbation theory using the Fermi golden rule. This transfer is assumed to be elastic, i.e., the energy of the electron is assumed to be the same before and after the transfer.

Let us calculate the transition rate  $\gamma_1^+$  for one electron from electrode 1 the molecule. The molecule begins in the state  $N$  and goes into the state  $N + 1$  after the transfer. Assuming a constant density of states  $\rho_1$  for the electrode,<sup>4</sup> about its electrochemical potential  $\mu_1$ , and denoting the element of the coupling matrix between the two states of the molecule by  $T$ , we obtain

$$\gamma_1^+ = t_1 f(\mu_{\text{mol}}(N + 1), \mu_1), \quad t_1 = \frac{2\pi}{\hbar} |T|^2 \rho_1.$$

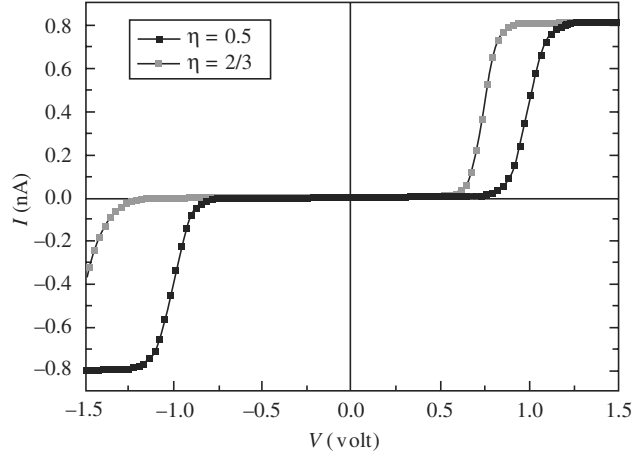
In the same way, the transition rates  $\gamma_1^-$  from the molecule towards electrode 1, and  $\gamma_2^+$  (resp.  $\gamma_2^-$ ) from electrode 2 (resp. the molecule) towards the molecule (resp. electrode 2) are given by

$$\gamma_1^- = t_1 \left[ 1 - f(\mu_{\text{mol}}(N + 1), \mu_1) \right],$$

$$\gamma_2^+ = t_2 f(\mu_{\text{mol}}(N + 1), \mu_2), \quad \gamma_2^- = t_2 \left[ 1 - f(\mu_{\text{mol}}(N + 1), \mu_2) \right],$$

where

<sup>4</sup> This is a good approximation for gold, the metal generally used for connections.



**Fig. 13.15.**  $I(V)$  characteristics calculated in the weak coupling model with  $\mu_1 = \mu_2 = 3.3 \times 10^{-6}$ . Electron affinity  $A = 0.5$  eV. The broadening of the steps is due to the temperature  $k_B T = 0.025$  eV. The observed asymmetry is caused by the coupling asymmetry

$$t_2 = \frac{2\pi}{\hbar} |T|^2 \rho_2 .$$

The current circulating through the molecule is now obtained by counting the electrons transferred from electrode 1 to electrode 2 and those transferred from electrode 2 to electrode 1. The rate of transfer of electrons from electrode 1 to electrode 2 is  $\gamma_1^+ \gamma_2^- / \gamma_\Sigma$ , whilst that from electrode 2 to electrode 1 is  $\gamma_2^+ \gamma_1^- / \gamma_\Sigma$ , with  $\gamma_\Sigma = \gamma_1^+ + \gamma_1^- + \gamma_2^+ + \gamma_2^-$ . The current is therefore given by

$$I = e \left( \frac{\gamma_1^+ \gamma_2^-}{\gamma_\Sigma} - \frac{\gamma_2^+ \gamma_1^-}{\gamma_\Sigma} \right) = e \frac{t_1 t_2}{t_1 + t_2} \left[ f(\mu_{\text{mol}}, \mu_1) - f(\mu_{\text{mol}}, \mu_2) \right] .$$

Note that the expression obtained here for the current is exactly the same as for the simple model of a discrete state calculated in the strong coupling case, without broadening but with a coupling given by<sup>5</sup>

$$\Gamma_i = \pi |T|^2 \rho_i .$$

Figure 13.15 shows the  $I(V)$  characteristics calculated using the last formula. The position of the step is associated with the electron affinity  $A$  divided by  $\eta$  for positive biases and  $A/(1 - \eta)$  for negative biases.

<sup>5</sup> This similarity may lead to confusion. It should be remembered that, in the strong coupling case, it is the energy of the molecular orbitals which determines the current, whereas in the weak coupling case, it is the electron affinity  $\mu_{\text{mol}}(N + 1)$  for a one-level model and the terms  $\mu_{\text{mol}}(N + i)$  (energy cost for adding a number  $i \in \mathbb{Z}$  of electrons) for a multilevel model.

This simple model illustrates the main principles of transport in the weak coupling case. We shall not go into the details of how this system can be made to operate as a single-electron transistor (see Chap. 11).<sup>6</sup> The models used to account for experimental results are complemented by taking into account several levels, excitations, and interlevel transitions [51, 52].

### 13.3.3 Functions

In this section, we shall describe several functions that can be carried out with electrically contacted molecules.

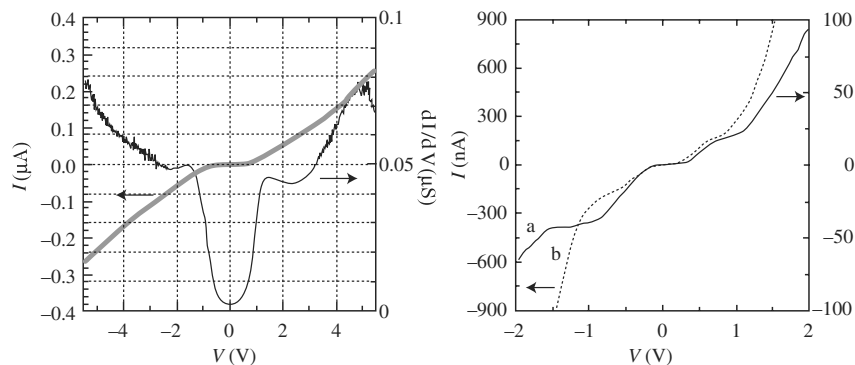
#### Molecular Wires

Unlike the aliphatic chains, in which all chemical bonds are saturated and which thus behave as insulators, the small  $\pi$ -conjugated oligomers are considered as prototypes for conducting molecular wires. The smaller HOMO–LUMO gap in these oligomers, around 3 eV compared with 8–9 eV in the aliphatic chains, explains why they are more efficient for electron transport. This greater efficiency also turns up in non-resonant tunnel transport characteristics. This regime corresponds to the strong coupling scenario described above, but out of resonance. The transport models discussed above can be used to express the conductance  $G = G_0 e^{-\beta L}$  of the metal–molecule–metal junction, where  $G_0$  is the contact conductance,  $\beta$  the tunneling decay factor, and  $L$  the length of the molecule [53]. In  $\pi$ -conjugated oligomers, the tunneling decay factor  $\beta$  is of the order of 0.2–0.6  $\text{\AA}^{-1}$  compared with 0.6–1.0  $\text{\AA}^{-1}$  for the aliphatic chains [53]. In other words, electron transport is possible over a greater distance in  $\pi$ -conjugated oligomers. The main oligomers studied have been the phenyl and thiophene oligomers, the oligophenylene vinylenes, the oligophenylene ethylenes, and the carotenoids. The number of monomers is generally between one and four, implying maximal lengths of the order of 25  $\text{\AA}$ . These molecules always have one functionalised end, e.g., by a thiol –SH, allowing them to attach themselves by chemisorption onto a gold electrode. Sometimes both ends are functionalised so that they can attach to both electrodes. The transport properties have been studied with various electrode configurations (see Sect. 13.3.1), STM or C-AFM and break junctions, where a small number of molecules are measured, or with monolayers sandwiched between two electrodes (crossed metal wires or nanopores).

The current–voltage curves are highly nonlinear, with steps (peaks in the derivative curve  $\partial I/\partial V$ ) [21, 23, 24]. Figure 13.16 shows two typical cases, for a benzene molecule and a terthiophene molecule [21, 23]. These steps correspond

<sup>6</sup> Formally, the effect of the electrostatic potential applied via a gate simply amounts to varying the chemical potential of the electrodes in the framework used above, adding a term  $-eC_g/CV_g$ , where  $V_g$  is the voltage applied to the gate.

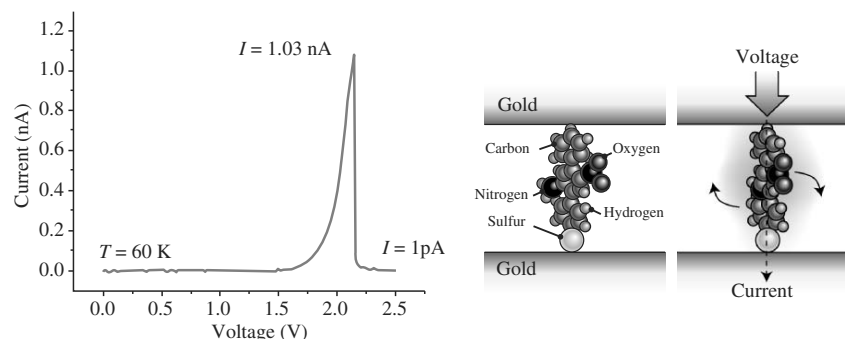




**Fig. 13.16.** Two examples of  $I$ - $V$  curves for  $\pi$ -conjugated oligomers inserted into a break junction. *Left:* benzene dithiol [21]. *Right:* terthiophene dithiol [23]

to a charge transfer between the electrodes via the molecular orbitals in resonance between the Fermi levels of the two electrodes (see Sect. 13.3.2). The measured conductance, or the current amplitude (taken at a given voltage and evaluated per molecule), depends on several factors. Apart from the intrinsic conductance of the molecule, the coupling between the molecule and the electrodes is a crucial factor. Likewise, between measurements made on an isolated molecule and those made on a collection of molecules, e.g., a monolayer, intermolecular interactions can modify the efficiency of electron transport. These interactions broaden the molecular orbitals and are therefore likely to enhance transport. In a recent review of around twenty results in this area, the observed currents thus varied from  $10$  to  $10^4$  pA/molecule [54]. The role of the chemical bond between the molecule and the electrode is clearly visible: the current increases by a factor of 10 when we compare the same oligophenylene ethylene connected chemically (via an S-Au bond) to both electrodes instead of to just one [55]. Likewise, the nature of the chemical bond is relevant, because it modifies the contact conductance  $G_0$ . In agreement with theoretical predictions [56, 57], it has been observed that electron transport is more efficient when a terthiophene is attached to a gold electrode by a selenium atom rather than by a sulfur atom [17, 46].

A last interesting feature of electron transport through these  $\pi$ -conjugated oligomers concerns the effects of conformation. The conductance of the molecule is reduced when the  $\pi$  groups are perpendicular to each other, whilst a coplanar conformation allows a good coupling of the  $\pi$  orbitals and thereby favours charge transfer along the molecule. A possible example [32, 33] of these conformation effects on the transport is the significant decrease in the current observed beyond a certain threshold voltage (negative differential resistance, see Fig. 13.17), which would correspond to a change in the redox state of the molecule, the central  $\pi$  group being substituted by an amine group  $\text{NH}_2$  and a nitro group  $\text{NO}_2$ , and hence to a change in its conformation. Recent

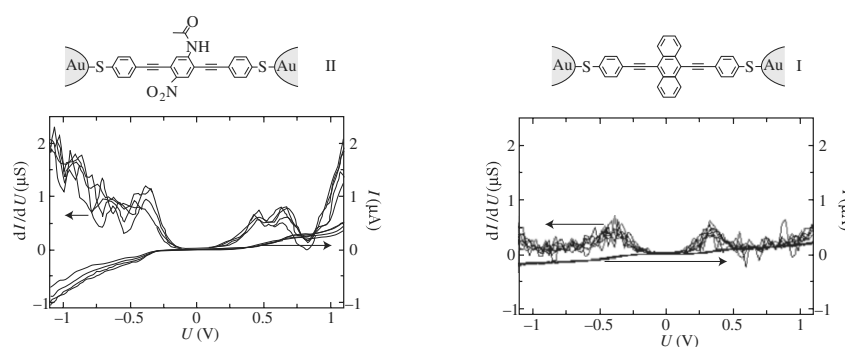


**Fig. 13.17.** *Left:*  $I$ - $V$  curve (at 60 K) of an amino-di(ethynylphenyl)-nitrobenzenethiol monolayer, with a nanopore configuration for the contacts (see Sect. 13.3.1). *Right:* Schematic representation of the assumed conformations in the conducting and insulating states of the molecule [32]

theoretical work [58] confirms this analysis. Another example [24] is provided by the observation, on the scale of a small number of molecules in a break junction, of the correlation between the conformational symmetry/asymmetry of the molecule and the bias symmetry/asymmetry of the respective  $I$ - $V$  curves (see Fig. 13.18).

## Diodes

Just as semiconductor electronics began with the invention of the  $p$ - $n$  junction, the first scientists to investigate molecular electronics sought to invent a molecular analogy of this current rectifying diode [1]. The rectifying diode is indeed a very simple component. However, in combination with simple resistances, one can already build logic circuits (diode-resistance logic). A similar



**Fig. 13.18.** Measurements made using the break junction technique [24]. *Left:* Asymmetric  $I$ - $V$  curve obtained with the asymmetric molecule (II). *Right:* Symmetric  $I$ - $V$  curve obtained with the symmetric molecule (I)

possibility also exists for negative differential resistance (NDR) diodes like those mentioned just above [32] on molecular wires. We shall not go into further detail concerning these resonant tunneling devices (RTD).

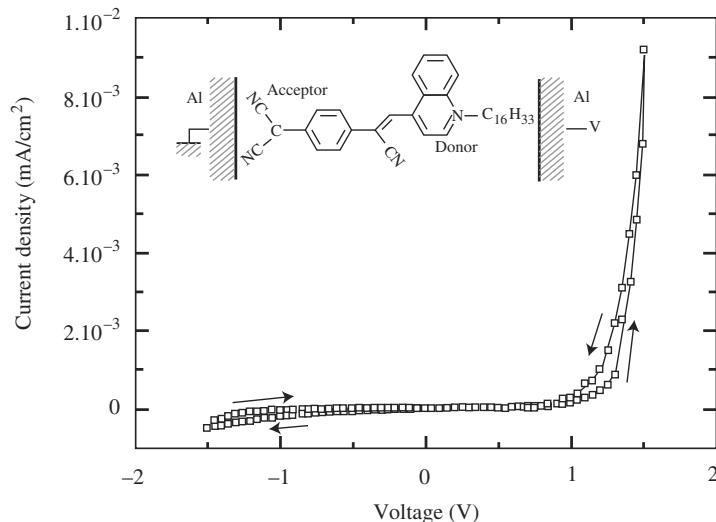
Thirty years after the theoretical proposal of Aviram and Ratner [1], we must ask what progress has been made in understanding the finer points of current rectifying molecular diodes and how far we have actually got towards making them.

At the present time, the donor–bridge–acceptor-type molecule that has produced the most significant experimental results is  $C_{16}H_{33}$ -Q-3CNQ, the hexadecyl-quinolinium tricyanoquinodimethanide molecule (see the insert of Fig. 13.19). A Langmuir–Blodgett monolayer of this molecule sandwiched between two metal (Al or Au) electrodes does indeed produce a current rectifying effect (rectification ratio of about 30), as illustrated in Fig. 13.19 [59–63]. The direction of rectification [more current for a positive bias on the electrode located on the donor (quinolinium) side] corresponds to the suggestion of Aviram and Ratner that an intramolecular electron transfer is easier from the acceptor group towards the donor than conversely. However, there are two differences between the original idea of Aviram and Ratner and what actually happens with the molecule  $C_{16}H_{33}$ -Q-3CNQ:

- The bridge between the two groups is  $\pi$  conjugated in the experimental case, whereas it is composed of saturated ( $\sigma$ ) bonds in the theoretical proposal.
- The experimental molecule is substituted at the donor end by a long aliphatic chain. This chain is required to make the molecule amphiphilic and suitable for forming a monolayer by the Langmuir–Blodgett technique.

These two differences have an important impact on the electron transport properties, as we shall see.

With the  $\sigma$  bridge, the Aviram–Ratner model assumes that the molecular orbitals of the donor and acceptor groups remain localised. In reality, with the  $\pi$ -conjugated bridge, this is not at all the case, as is shown by the theoretical calculation using the density functional theory (DFT). The HOMO and LUMO are delocalised over the whole of the two groups (see Fig. 13.20) [64]. A direct consequence of this result is that the current–voltage characteristic (calculated by the self-consistent tight-binding method) of a monolayer of the molecule Q-3CNQ (the D–A system without the aliphatic chain) will be symmetric (see Fig. 13.21a), and hence will have no rectifying effect on the current. The same calculation with the aliphatic chain yields an asymmetric characteristic (see Fig. 13.21b), in qualitative agreement with experiment [64]. In this case, it is the introduction of a geometrical asymmetry into the metal–molecule–metal structure which is responsible for this current rectifying effect. Indeed, the D–A group is better coupled electrically with one electrode than with the other, being separated from the latter by an aliphatic chain that serves as an electrical insulator. The electrical potential is not distributed uniformly over the whole molecule and the D–A group is only therefore subject



**Fig. 13.19.** Typical  $I(V)$  curve for the metal/ $C_{16}H_{33}$ -Q-3CNQ/metal junction shown in the *insert*. The electrode on the acceptor side is earthed and a bias is applied to the electrode on the donor side [60]

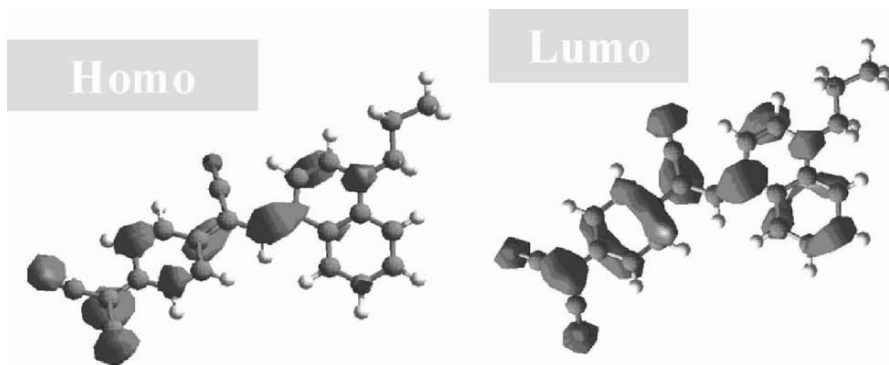
to a fraction  $\eta$  of the applied bias (see Sect. 13.3.2). To a first approximation, this fraction is given by

$$\eta = \frac{1}{2 \left( 1 + \frac{\varepsilon_{\pi} d_{\sigma}}{\varepsilon_{\sigma} d_{\pi}} \right)},$$

where  $\varepsilon_{\sigma}$ ,  $\varepsilon_{\pi}$ ,  $d_{\sigma}$ , and  $d_{\pi}$  are the dielectric constants and thicknesses of the  $\pi$ -conjugated part and the aliphatic chain of the  $C_{16}H_{33}$ -Q-3CNQ monolayer, respectively.

As a general rule, an asymmetry in the coupling of the molecule with the two electrodes is liable to lead to such a result (see Sect. 13.3.2). In the case of the diode based on  $C_{16}H_{33}$ -Q-3CNQ molecules, since the energy difference  $E_F - E_H$  between the Fermi level of the metal and the HOMO is smaller than the difference  $E_L - E_F$  with respect to the LUMO, it is easier to obtain resonance between the HOMO and the Fermi level of the metal for a positive bias  $V^+$  such that  $\eta eV^+ \approx E_F - E_H$ . A negative bias with larger absolute value would be needed for resonance with the LUMO. It is these two effects, geometric and energetic asymmetries, which induce the rectifying behaviour of the metal/ $C_{16}H_{33}$ -Q-3CNQ/metal junction.

These asymmetry mechanisms can be generalised [65] and can be exploited to fabricate molecular diodes with other  $\pi$ -conjugated groups [48], and in particular, with groups that are easier to synthesise than the  $C_{16}H_{33}$ -Q-3CNQ molecule. Figure 13.22 illustrates for a self-assembled organic layer (see



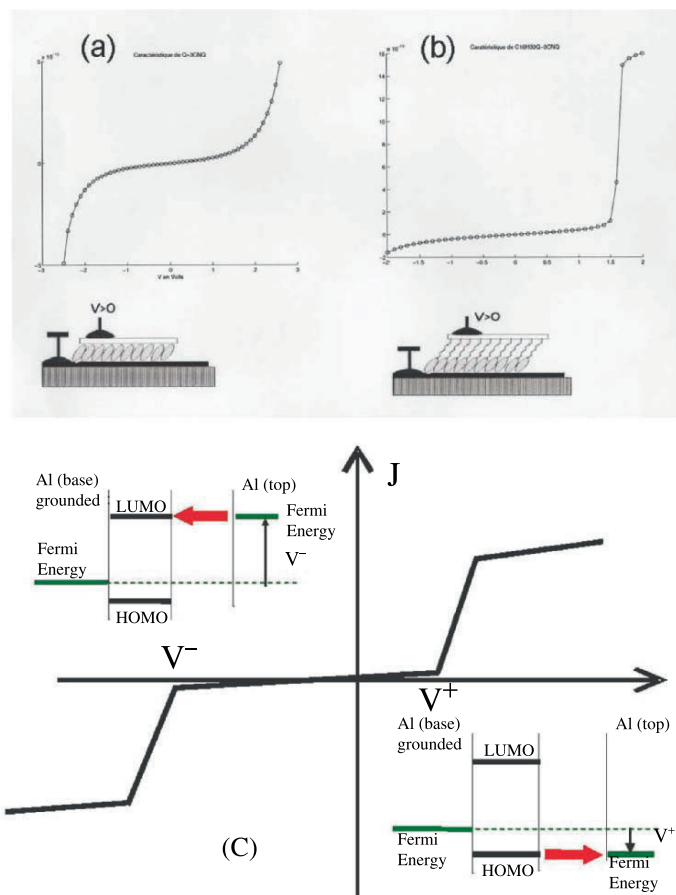
**Fig. 13.20.** Representation of the density of states of the HOMO and LUMO calculated by DFT. To simplify the calculations, a short chain (propyl) was used, the  $\pi$  orbitals being largely independent of the length of this chain [68]

Sect. 13.5.1) on a highly  $n$ -doped (degenerate) silicon substrate. The monolayer is synthesised by first chemically grafting alkyltrichlorosilane chains, then grafting the  $\pi$ -conjugated function onto this first monolayer [66, 67]. A good current rectification effect is observed here too, but for negative biases, the  $\pi$ -conjugated group being above the chain (in contrast to what happens with  $C_{16}H_{33}Q-3CNQ$ ). Rectification ratios of about 35 and rectification threshold biases ranging from  $V_T = -0.3$  to  $-0.9$  V are recorded. Molecular diodes of this type have thus been made with simple donor  $\pi$ -conjugated groups such as phenyl, thiophene, pyrene, anthracene, etc., which are easy to synthesise and/or commercially available, and for different lengths of the alkyl chain, i.e., from 6 to 15  $CH_2$  groups.

It only remains now to optimise these molecular systems in such a way as to adjust the rectification threshold voltage by judicious choice of the  $\pi$ -conjugated group and the length of the alkyl chain, and to further increase the rectification ratio.

### Electromechanical Components

As shown in Sect. 13.3.2, there is a relationship between molecular structure and transport properties. Consequently, any change in the conformation induced by an external stimulus is likely to change the electronic levels of the molecule and hence also its transport properties within a device. In this way, a mechanical stress can be used to deform a molecule and thereby produce an electromechanical component. This has been demonstrated in the configuration illustrated in Fig. 13.23. An STM was used to measure the transport properties of a  $C_{60}$  molecule, whilst the STM tip simultaneously deformed the molecule to varying degrees [69]. It was thus shown that a molecule in the described configuration could function as an amplifier. As a typical example,

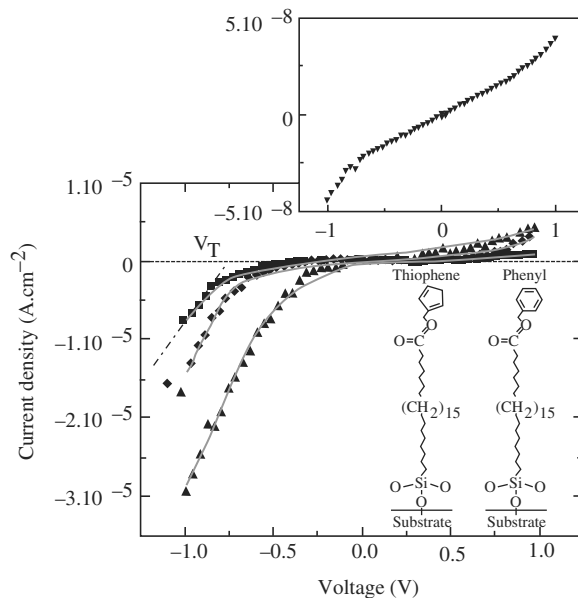


**Fig. 13.21.** Calculated  $I(V)$  curves (a) for a metal/Q-3CNQ/metal junction, (b) for a metal/C<sub>16</sub>H<sub>33</sub>-Q-3CNQ/metal junction. (c) Simplified energy representation when the HOMO and LUMO levels resonate for the two polarities [68]

a modulation of 20 mV applied to the piezoceramic of the STM led to a modulation of 100 mV in the output voltage measured across a load resistance.

For small compressions, the current varied exponentially with the compression. A deformation of 0.1 nm in the C<sub>60</sub> molecule led to a current variation by two orders of magnitude. Part of the gain in this component can be attributed to the electromechanical conversion of the electrical signal applied to the piezoceramic. The rest originates in a modification of the molecular orbitals of the C<sub>60</sub> molecule, due to electronic repulsion between the orbitals under stress.

Finally, note that similar effects have also been observed using the break junction technique to contact bithiolterthiophene molecules. In this case, a



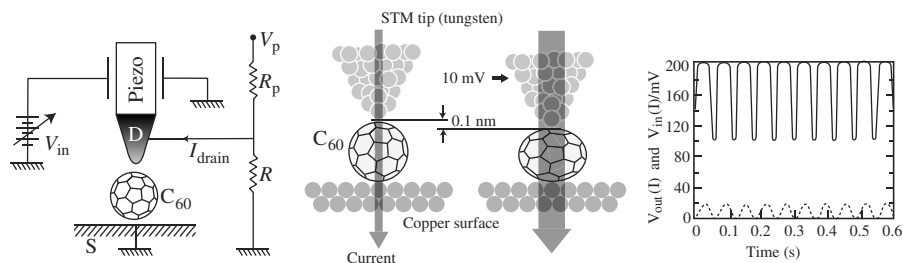
**Fig. 13.22.** Current–voltage curves for three  $\text{Si-}n^+/\sigma\text{-}\pi$  monolayer/Al junctions: thiophene–short chain  $(\text{CH}_2)_6$  ( $\blacktriangle$ ), phenyl–long chain  $(\text{CH}_2)_{15}$  ( $\blacklozenge$ ), and thiophene–long chain  $(\text{CH}_2)_{15}$  ( $\blacksquare$ ). The threshold bias for rectification ( $V_T$ ) is indicated in one case. *Insert:* Molecular structure of two of the long-chain junctions. The upper current–voltage curve belongs to a reference junction without the  $\pi$ -conjugated group  $\text{Si-}n^+/\text{alkyl chain}/\text{Al}$ , showing that there is not then any current rectification effect [48]

change of just  $0.3 \text{ \AA}$  is enough to significantly modify the transport properties [23].

### Molecular Bistables and Memories

The catenanes and rotaxanes are two classes of molecules which exhibit bistable behaviour. These molecules generally comprise two interlocked but relatively mobile parts, one around or one inside the other, e.g., a ring around a rod, two interlocking rings, etc. (see Fig. 13.24) [70–73]. These molecules can adopt different conformations depending on their redox state. A change in the redox state, e.g., under the effect of an optical or chemical excitation, can trigger the displacement of one of the mobile elements in such a way as to minimise the total energy of the molecule.

This type of molecule can be used to make molecular memory cells in which each bistable state corresponds to the coding of a piece of binary information, the 0 or 1 of Boolean logic. An electrical bistable effect has been observed with metal–molecule–metal junctions (see Fig. 13.24), where a monolayer of



**Fig. 13.23.** Electromechanical amplifier based on  $C_{60}$ . The *center figure* shows the principle of operation in a schematic manner. The current intensity is represented by the width of the *vertical arrow*. *On the right*, the output voltage measured on the resistance  $R$  is plotted as a function of the input voltage applied to the STM piezoceramic [69]

these molecules mixed with phospholipid acids is sandwiched between two electrodes [74–76]. A voltage pulse of 1.5–2 V switches from the off state to the on state (writing the memory). The information can then be read by measuring the current at low voltage, e.g., 0.5 V. The system is returned to the off state by applying an opposite pulse of  $-1.5$  to  $-2$  V (erasure). A prototype with 64 non-volatile memory cells has been made in a crossbar architecture using these components, with performance in terms of the current ratio in the two states  $I_{\text{on}}/I_{\text{off}} \approx 10$ –50, data retention time  $\sim 24$  hr, and endurance, i.e., number of read/write cycles  $\approx 100$  (see Fig. 13.24) [77, 78]. The metallic rows and columns in this memory plane have a width of 40 nm and are made by a cheap technique using nanoimprint lithography (see Chap. 1). The potential advantages of this molecular memory architecture are:

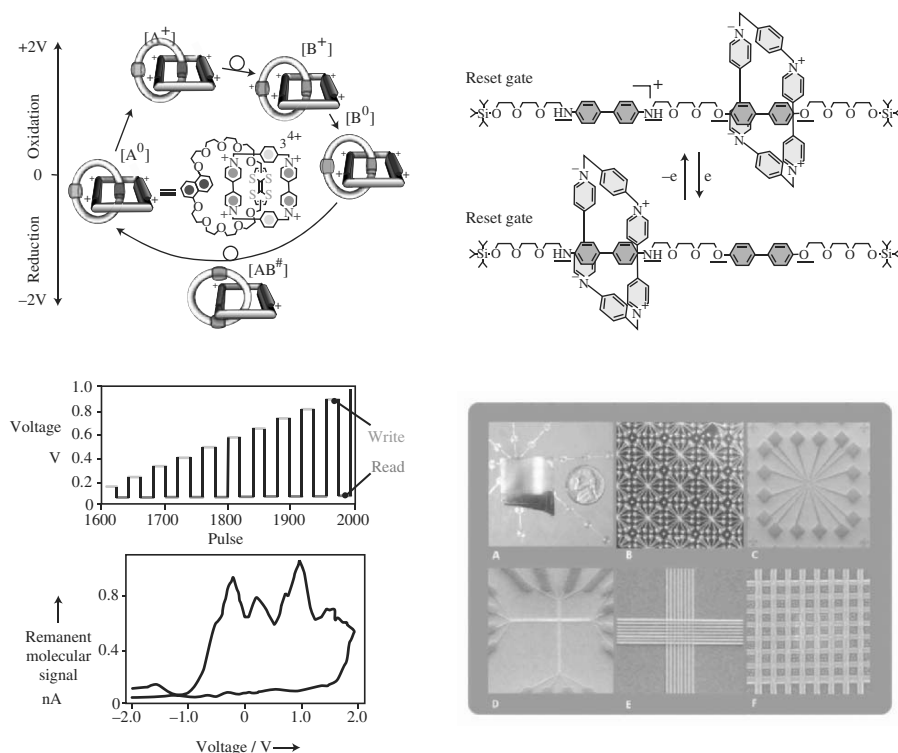
- low cost of fabrication,
- high integration density (6.4 Gbit/cm<sup>2</sup> in the above example, excluding peripheral addressing circuits),
- defect-tolerant architecture,
- easy post-processing above traditional CMOS circuits (hybrid nanoelectronics).

### Molecular Transistors

This term encompasses three-electrode devices in which two electrodes, the source and drain, are used to contact the molecule, whilst the third, the gate, insulated from the molecule by an insulating layer, serves to influence transport within the molecule in an electrostatic manner (see Fig. 13.2d). Depending on the type of coupling between the molecule and the metal, the mechanism whereby the gate influences transport can be completely different (see Sect. 13.3.2).

In the strong coupling regime, the electrostatic influence operates by modifying the molecular orbitals via the transverse electric field (generated by the electrostatic potential applied by the gate). In the context of the discussion





**Fig. 13.24.** Examples of a catenane (*top left*) and a rotaxane (*top right*) [72,74,75] and a bistable  $I(V)$  curve (*bottom left*) obtained with the catenane molecule. *Bottom right*:  $8 \times 8$  memory cell using these molecular components. The image F shows rows and columns of width 40 nm in the plane of the cell. The molecular monolayer is inserted between the electrodes at each crossing point of a row and a column [77, 78]

in Sect. 13.3.2, this amounts to changing either the broadening or the position of the energy levels arising from the coupling of the molecular orbitals and the electrodes. No experimental confirmation of this mechanism has yet been obtained. This may be due to the very high order of magnitude of the electric field (around  $1 \text{ V}/\text{\AA}$ ) required to produce measurable effects (see the calculations by Di Ventura et al. [79]).

In the weak coupling regime, the electrostatic influence operates, as explained in Sect. 13.3.2, by changing the total energy of the molecule (or, in the reference system used in Sect. 13.3.2, by adding a term  $-eC_g/CV_g$  to the chemical potential of the electrodes, where  $V_g$  is the gate potential). An experimental confirmation of this effect has been obtained by McEuen and coworkers by measuring a  $C_{60}$  molecule in a nanogap obtained by electromigration (see Sect. 13.3.1 and Fig. 13.5). More recently, other (organometallic) molecules have been tested and shown to have a spin-dependent behaviour.

It should nevertheless be noted that an electrostatic gate is globally rather inefficient when it comes to influencing the potential perceived by a single molecule within a nanogap. This will in all likelihood restrict the use of molecular transistors to the role of a research tool. Indeed, given the size of the nanogap, screening by the source and drain electrodes is highly efficient and the influence of the gate decreases exponentially with the ratio between the height of the molecule (above the insulator forming the floor of the nanogap) and the width of the nanogap.

## 13.4 Components Based on Nanotubes

Over the last few years, the potential of carbon nanotubes (CNT) for nano-electronics has been confirmed by the fabrication of ever more complex devices based on these macromolecules, which can exist in metallic or semiconducting form, as discussed in Chap. 8.

### 13.4.1 Field-Effect Transistors

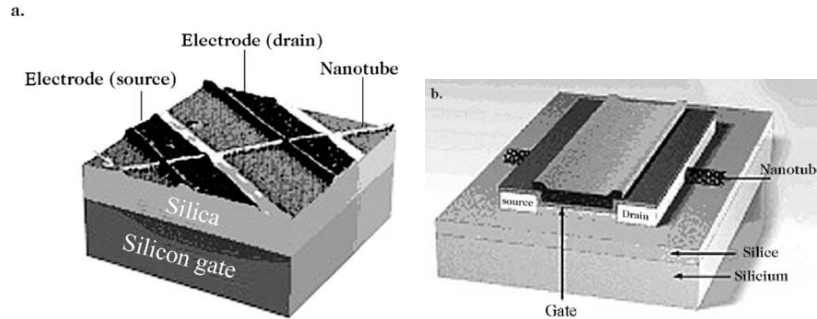
Among the components that have been made, one of the most significant is the carbon nanotube field-effect transistor (CNTFET) in 1998. The structure of these first CNTFETs is very simple. A single-walled nanotube (SWNT) is positioned in such a way as to connect two electrodes which play the roles of source and drain, as shown in Fig. 13.25a. Using a basic idea analogous to the MOS transistor, the passage of the current from the source electrode to the drain electrode is modulated by the field applied via the gate electrode. The latter is insulated from the semiconducting nanotube, which forms the transistor channel, by a thick silica insulator [80, 81].

In order to make a CNTFET, a semiconducting nanotube must be inserted between the source and drain electrodes. Various methods can be used (see Sect. 13.5.1). The nanotube can be grown directly in situ on the substrate or it can be placed there after growth. Note that in neither case does there yet exist a method for deciding in advance whether the nanotube will be semiconducting or metallic.<sup>7</sup> The nature of the nanotube must therefore be ascertained by subsequent tests.

In the first CNTFET devices, the metallic electrodes were made on silica ( $\text{SiO}_2$ ) obtained by thermal growth on the silicon wafer. The latter then also served as the gate and this arrangement was known as a back-gate configuration. In fact, the source and drain electrodes can be fabricated either before or after placing the nanotube, but better contacts have been observed experimentally in the latter case. Finally, different geometries have also been

---

<sup>7</sup> However, in 2003, certain rather promising ideas were demonstrated for sorting nanotubes in solution.



**Fig. 13.25.** Nanotube transistors. (a) The substrate serves as the gate. Courtesy of C. Dekker. (b) The gate is deposited on the nanotube. Courtesy of P. Avouris

proposed for the gate electrode: back-gate arrangement in Fig. 13.25a and top-gate arrangement in Fig. 13.25b.

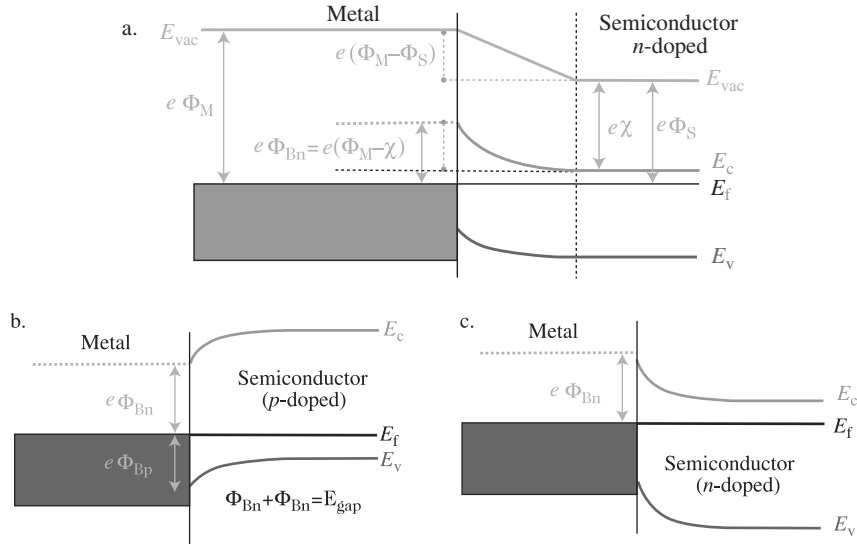
Independently of the fabrication arrangements devised for CNTFETs, three basic setups can be envisaged. The first assumes that the  $p$ -CNTFET works like a depletion transistor [89]. In this case, the nanotube is considered to be uniformly doped and ohmic contacts are assumed at each end. The second possibility is a MOSFET-type operation in which the part without a gate is highly doped [89]. The third alternative is a device analogous to a Schottky barrier field-effect transistor (SBFET) [90].

Experimental results obtained so far are unable to clarify exactly how CNTFET devices work, or even whether the devices realised are all the same. However, it would seem that the majority of these results can be conveniently explained by a simple SBFET-type description. Indeed, it was originally assumed that the applied gate voltage modified the conductance of the nanotube as in an ordinary FET. But a theoretical analysis of the work functions of the relevant materials suggests that there are significant Schottky-type barriers at metal–nanotube junctions [91, 92], whilst more recently experimental results have shown that these barriers at the contacts dominate the behaviour and performance of CNTFETs.

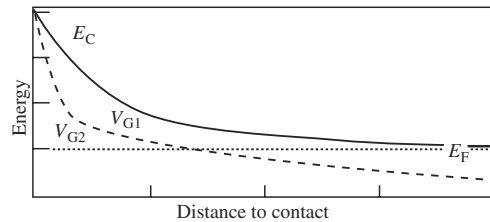
Recal that a metal– $n$ -doped semiconductor junction can be represented schematically as in Fig. 13.26a [89]. The height  $e\Phi_{Bn}$  of the Schottky barrier of this junction depends on the work function  $e\Phi_M$  of the metal and the electron affinity  $\chi$  of the semiconductor. For a more detailed discussion, the reader is referred to [91].

In this model, the extent of band bending in the neighbourhood of the metal–semiconductor junction is related to the height of the Schottky barrier ( $e\Phi_{Bn}$  for electrons and  $e\Phi_{Bp}$  for holes) and the position of the Fermi level  $E_F$  of the isolated semiconductor, a function of its doping state (see Fig. 13.26b for  $p$ -type doping and Fig. 13.26c for  $n$ -type doping).

It has been shown that the electrical behaviour of transistors based on nanotubes depends mainly on the modulation of the barrier characteristics



**Fig. 13.26.** (a) Operating principle of metal–semiconductor junction. The height  $e\Phi_{Bn}$  of the Schottky barrier of this junction depends on the work function  $e\Phi_M$  of the metal and the electron affinity  $\chi$  of the semiconductor. (b) and (c): Heights  $e\Phi_{Bp}$  and  $e\Phi_{Bn}$  of the Schottky barrier for holes and electrons, respectively, for a  $p$ -doped or  $n$ -doped nanotube



**Fig. 13.27.** Conduction band bending at the source contact as a function of the distance to the contact for two gate voltages  $V_{G1} < V_{G2}$ . Adapted from [90]

under the effects of the gate voltage, rather than on the modulation of the channel (nanotube) conductance. This can be explained by means of the diagram in Fig. 13.27, where the behaviour of the conduction band with the contact distance (nanotube–metal junction) is plotted for two different values of the gate voltage  $V_G$ .

The width of the Schottky barrier can be modified via the gate voltage. For example, it is reduced when the gate voltage changes from  $V_{G1}$  to  $V_{G2} > V_{G1}$  (see Fig. 13.27). In this case, the probability of the electrons tunneling through the barrier is increased and the electrical behaviour of the transistor is thereby affected (more current goes through).

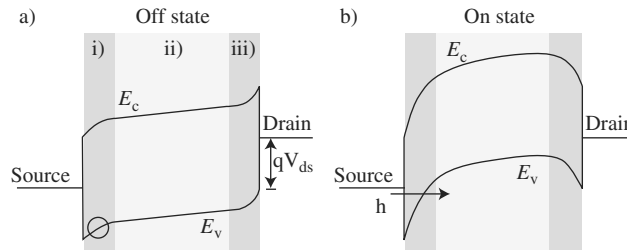
Figure 13.28 is a qualitative representation of the conduction and valence band profiles in a nanotube transistor (*p*-type CNTFET, hole transport) for two values of the gate voltage  $V_G$ . Regions (i) and (iii) represent segments of the nanotube located near the collecting (drain) electrode and the emitting (source) electrode (Schottky barrier regions), while region (ii) is the central part of the nanotube.

When the gate voltage is less than the threshold voltage  $V_{th}$  (Fig. 13.28a), the current  $I_d$  between the emitter and collector is blocked by the barrier in the emitting region (i). In this case, the barrier width can be modulated by the gate, but also by the field between the emitter and the collector. The bulk part of the nanotube plays no role whatever in controlling the off-state current. This hypothesis has been confirmed by studies of the transport properties as a function of temperature. For a given value  $V_{ds}$  of the drain–source voltage, the current  $I_d$  increases exponentially with the gate voltage  $V_G$  due to the narrowing of the Schottky barrier in (i) (see Fig. 13.28a and the region marked ‘a’ in Fig. 13.29).

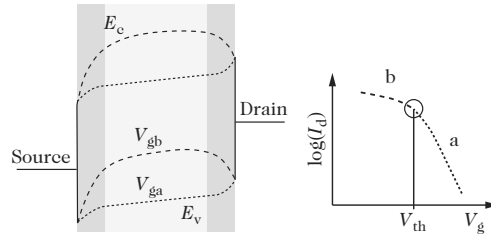
The behaviour of the transistor changes when the gate voltage is such that the valence band of the bulk region of the nanotube lines up with the Fermi level of the source electrode. This value of  $V_G$  determines the threshold voltage  $V_{th}$ . Beyond this threshold value, charge accumulates in the nanotube, accompanied by a lesser dependence of the band bending on the gate voltage. Consequently, the width of the Schottky barrier varies to a lesser degree and the tunnel current increases more slowly (see Fig. 13.28b and the region marked ‘b’ in Fig. 13.29).

For a fixed value of  $V_G$  greater than the threshold value  $V_{th}$ , the effect of the voltage applied between the source and drain can be understood from the diagram shown in Fig. 13.30: on the left, the evolution of the conduction and valence band profiles with changing  $V_{ds}$ ; on the right, the corresponding evolution of the current  $I_d$ .

It is easy to understand that the threshold voltage should depend on the position of the Fermi level of the bulk region of the nanotube and that  $V_{th}$  should be shifted when the tube is doped (see also Figs. 13.26b and c or Fig. 13.31b for the band diagram, and Fig. 13.32b for a numerical simulation



**Fig. 13.28.** Schematic picture of the transistor bands [93]. (a) Off state  $V_G < V_{th}$ . (b) On state  $V_G > V_{th}$ .



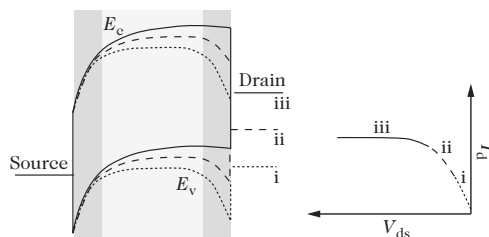
**Fig. 13.29.** Band diagram of the transistor for two gate voltages, (a) below and (b) above the threshold. Adapted from [94]

of the current). At the same time, the height of the Schottky barrier can be altered by changing the work function of the metal electrode (see also Figs. 13.26a and 13.31a). This leads to a modification of the symmetry of the  $I_d/V_G$  curves and a slight effect on the on–off operating range of the device, as depicted in Fig. 13.32a.

We end this discussion of CNTFETs by mentioning that the performance of these devices has been considerably improved over the past two years, since it has been understood how they function. Table 13.1 compares their performance with silicon MOSFET devices with comparable geometry. [All lateral dimensions of components are referred to  $1\ \mu\text{m}$  by assuming the parallel connection of ca.  $1000/d$  (nm) nanotubes, where  $d$  is the nanotube diameter.] The table clearly shows that CNTFET performance has reached remarkable levels today, enough to interest industrial manufacturers.

Using this type of transistor, simple logic circuits have been built, such as an inverter, a NOR, an SRAM, and a ring oscillator [82].

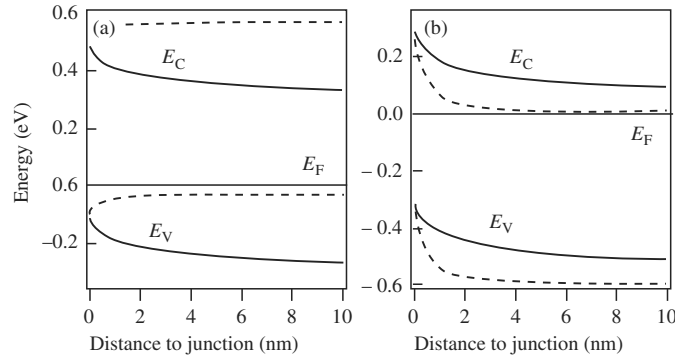
A standard method for building logic circuits by analogy with CMOS circuits would be direct inclusion of nanotube devices which use electrons ( $n$ -type CNTFET) and holes ( $p$ -type CNTFET) as carriers. However, the first transistors based on nanotubes invariably exhibited  $p$ -type characteristics. In order to make logic gates using these complementary devices, the  $p$ -CNTFETs must first be converted into  $n$ -CNTFETs.



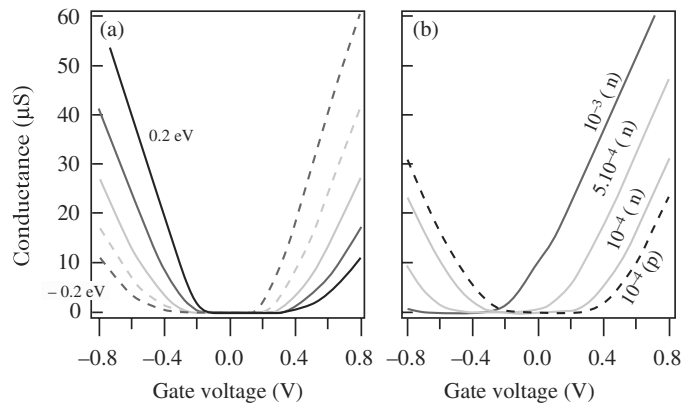
**Fig. 13.30.** Band diagram of the transistor for three different source–drain voltages and variation of the source–drain current as a function of the source–drain voltage. Adapted from [93]

**Table 13.1.** Comparison of the performance of CNTFET devices and three advanced silicon FET devices [95]

	<i>p</i> -CNT FET 1 $\mu\text{m}$ (1 V)	<i>p</i> -CNT FET 260 nm (1 V)	<i>p</i> -CNT FET 1.4 $\mu\text{m}$ (1 V)	<i>p</i> -CNT FET 3 $\mu\text{m}$ (1.2 V)	<i>p</i> -CNT FET 100 nm (1.5 V)	FinFET 10 nm (1.2 V)	MOSFET 14 nm (0.9 V)
On current [mA/ $\mu\text{m}$ ]	0.714	2.142	2.99	3.5	1.4 <i>n</i> -FET 0.46 <i>p</i> -FET	0.450 <i>n</i> -FET 0.360 <i>p</i> -FET	0.215 <i>p</i> -FET
Transconductance [ $\mu\text{S}/\mu\text{m}$ ]	214	2284	6666	6000	1000 <i>n</i> -FET 460 <i>p</i> -FET	500 <i>n</i> -FET 450 <i>p</i> -FET	360 <i>p</i> -FET
Subthreshold slope [mV/dec]	730	130	80	70	90	125 101	71
On resistance [ohm/ $\mu\text{m}$ ]	1400	660	360	342	1442 <i>n</i> -FET 3260 <i>p</i> -FET	2653 <i>n</i> -FET 3333 <i>p</i> -FET	4186 <i>p</i> -FET
Gate length [nm]	1030	260	1400	2000	130	10	14
Normalised gate length [1/nm]	4/150 = 0.026	4/15 = 0.26	80/1 = 80	25/8 = 3.12	4/2 = 2	4/1.7 = 2.35	4/1.2 = 3.33
Mobility [ $\text{cm}^2/\text{Vs}$ ]	68	–	1500	3000	–	–	–
Off current [nA/ $\mu\text{m}$ ]	7	7	–	1	3	10	100



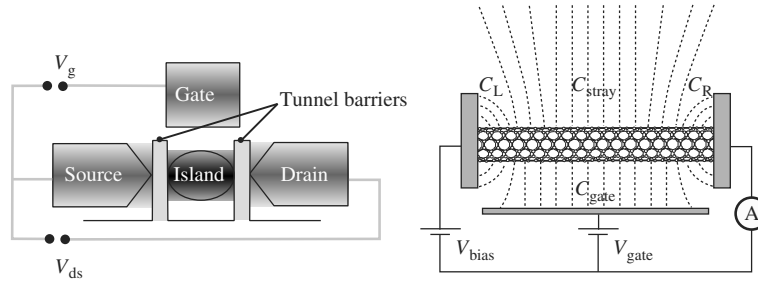
**Fig. 13.31.** Schematic view of the conduction and valence bands of the nanotube at the corresponding source contact. **(a)** For a metal whose work function has been increased by 0.2 V with respect to the case where the Fermi level of the metal corresponds to a midgap position of the nanotube. **(b)** Effect of  $n$ -doping the nanotube. In both diagrams, *continuous curves* correspond to zero gate voltage and *dashed curves* to a gate voltage of  $-0.5$  V [90]



**Fig. 13.32.** Conductance of the nanotube transistor channel as a function of the applied gate voltage. **(a)** Variation of the work functions of the source and drain electrodes from 0.2 eV to  $-0.2$  eV in steps of 0.1 eV. **(b)** Effect of  $n$ -doping by  $10^{-3}$ ,  $5 \times 10^{-4}$ , and  $10^{-4}$ , and  $p$ -doping by  $10^{-4}$  (atomic fractions) [90]

This conversion (from  $p$ -CNTFET to  $n$ -CNTFET) can be made by doping with an electropositive element such as potassium [83] or vacuum annealing [84].  $p$ - and  $n$ -CNTFET devices fabricated on the same substrate have been successively connected to build logic gates [84]. However, it should be pointed out that, even though CNTFET devices have been fabricated with comparable or better performance than Si MOSFET devices with comparable geometry [85–88], the technological aspects are still in their early stages.





**Fig. 13.33.** *Left:* Single-electron transistor. *Right:* Nanotube application. The capacitances are shown in the right-hand image. Tunnel barriers are formed at the contact between the nanotube and the electrodes

The structures of these devices are still primitive and certain aspects of their physical properties remain to be properly explored.

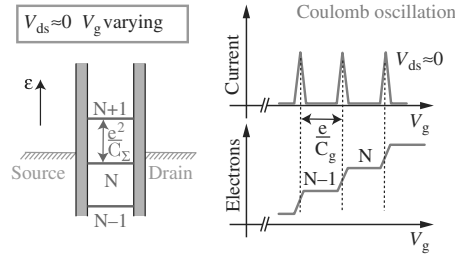
### 13.4.2 Single-Electron Transistors (SET)

Owing to their structure and their behaviour as coherent quantum conductors (see Chap. 8), carbon nanotubes are a natural candidate for making devices like single-electron transistors (SET) (see Chap. 11). The general setup and the nanotube device are shown in the left- and right-hand diagrams of Fig. 13.33, respectively.

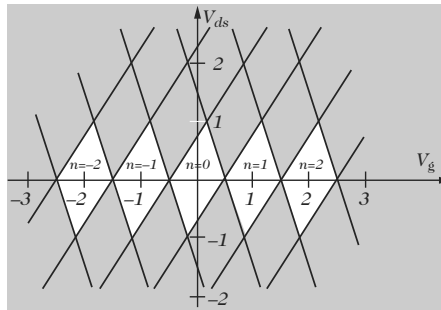
In this context, we will assume that the contacts of the nanotube behave as tunnel barriers. To obtain a current through the system, electrons must be injected into and collected from the output of the nanotube. In this process, all the capacitances connecting the nanotube to its surroundings (indicated by  $C_L$ ,  $C_R$ ,  $C_{\text{gate}}$ , and  $C_{\text{stray}}$ ) must be charged up and discharged. For each electron, the addition energy associated with this process is  $E_{\text{add}} = e^2/C_{\Sigma}$ , where  $C_{\Sigma} = C_L + C_R + C_{\text{gate}} + C_{\text{stray}}$ .

For a nanoscale conductor, i.e., in which the value of  $C_{\Sigma}$  is small, the energy associated with addition of one electron is very high compared with  $k_B T$ . It follows that the current is blocked in this regime (Coulomb blockade) and that the number of electrons in the nanotube is fixed, e.g., at  $n$ . This blockade can be removed by varying the electrostatic potential of the nanotube by means of the gate with which it is capacitively coupled. An electron can then enter and leave the nanotube and a current can therefore circulate. Further increasing the gate voltage leads in turn to another current blockade, with a fixed number  $n + 1$  of electrons in the nanotube. We thus observe a resonant peak in the conductance when the system evolves between two blockade regimes. The term ‘single-electron transistor’ is used because the conductance between two contacts can be controlled (either induced or inhibited) by means of a third electrode. Figure 13.34 provides a summary of what is happening.

The blockade can also be lifted by increasing the source–drain bias. The two electrodes (source and drain) are directly coupled to the nanotube by



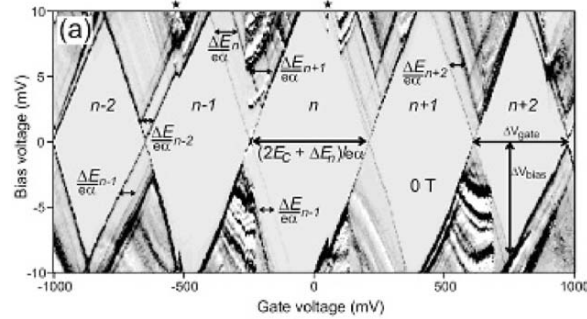
**Fig. 13.34.** Single-electron transistor running at low source–drain bias for different gate voltages



**Fig. 13.35.** Stability diagram in a single-electron transport configuration

tunneling contacts, but also capacitively. Measurements of the differential conductance as a function of drain–source voltage  $V_{ds}$  and gate voltage  $V_{gate}$  appear in the well known diamond arrangement in the  $(V_{ds}, V_{gate})$  plane, with vertex points  $(v_n \pm \Delta V_{gate}/2; 0)$  and  $(v_n \pm \delta V_{gate}; \pm \Delta V_{bias})$  [96]. This type of diagram is called the stability diagram (see Fig. 13.35). The parameter  $v_n$  is a constant and  $\delta V_{gate} < \Delta V_{gate}/2$  is an asymmetry parameter depending on the value of the various capacitances coupling the nanotube. Within a given diamond, the current is blocked and the number of electrons fixed. The width of the diamond is the distance  $\Delta V_{gate}$  between two of the conductance peaks mentioned above.

The first investigations using nanotubes were carried out by connecting a metallic nanotube in a transistor configuration and making measurements at low temperature. Note that the effects discussed here are only accessible if the metallic contacts on the nanotube are of tunnel type. These experiments showed that the low temperature transport properties of metallic nanotubes can be described by the Coulomb blockade theory, where the thermal energy  $k_B T$  is smaller than the charging energy  $E_C = e^2/2C_{tube}$  of the nanotube [97]. However, at very low temperatures, the energy gap  $\Delta E$  between two quantised levels of the nanotube can be resolved, i.e.,  $k_B T < \Delta E < E_C$ , in contrast with the classical regime at high temperatures where  $\Delta E < k_B T < E_C$ . In



**Fig. 13.36.** Stability diagram of a single-electron transistor based on a carbon nanotube [98]

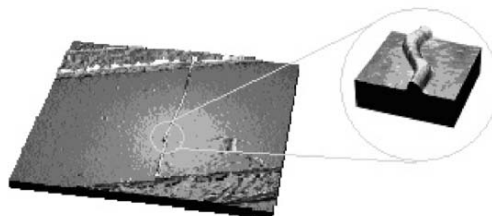
this case, when the thermal energy is less than the separation  $\Delta E$  between two quantum levels, the energy  $\Delta E$  must also be supplied in order to add an electron to the nanotube and the addition energy for electrons then becomes  $E_{\text{add}} = e^2/C_{\Sigma} + \Delta E$ .

Note that the energy gap between levels in the spectra of independent particles can be different for different occupations of the nanotube. This leads to a variation in the distance between conductance peaks

$$\Delta V_{\text{gate}} = \frac{2E_C + \Delta E_n}{e\delta},$$

where  $\delta = C_{\text{gate}}/C_{\text{tube}}$  is the capacitive coupling parameter of the nanotube and the gate and  $n$  is the number of electrons in the nanotube. We thus obtain the change in the addition energy as a function of the number of electrons, i.e. the addition spectrum.

The energy levels of the nanotube can also be found from the stability diagram. At the point where the Coulomb blockade has been lifted by a change in the source–drain bias, the occupation of the nanotube oscillates between  $n$  and  $n + 1$ . When the source–drain bias is increased, a higher energy level enters the voltage range for transport and hence contributes to transport. The increase in the probability of electron transfer causes a rise in the current and hence a peak in the differential conductance. However, it should be noted that the occupation number still oscillates between  $n$  and  $n + 1$ . Because the next energy level is higher in energy than the fundamental level already contributing to transport, it is called an excited state. If the capacitive couplings of this excited state with the source and drain electrodes and the gate electrode are equal to those of the ground state, its contribution appears as a line parallel to the side of the diamond. The distance (in gate voltage) between this line and the side of the diamond is equal to  $\Delta E/e\delta$ . A set of excited states of the nanotube will generate a set of extra lines. This set of lines is called the excitation spectrum. An example of the kind of measurements obtained for a metallic nanotube is shown in Fig. 13.36.



**Fig. 13.37.** AFM image of an SET operating at room temperature on the basis of a part of a nanotube insulated by tunnel barriers (created by deforming the nanotube), generated by AFM manipulation of the rest of the nanotube. Courtesy of C. Dekker et al.

Two important conclusions can be drawn from the various studies carried out on CNTSET devices. For one thing, the experiments confirm the quantum-coherent nature of the nanotube over lengths as long as several hundred nanometers. For another, the charging energy and level separation in nanotubes agree well with theoretical predictions deduced from band structure calculations for the nanotubes. The drawback with these measurements is that one is compelled to work at very low temperatures. To remedy this, one must reduce the size of the part of the nanotube serving as island.

Figure 13.37 shows an SET fabricated on a metallic nanotube by mechanical deformation of that nanotube using an AFM tip. The device comprises a short piece of nanotube (less than 20 nm long), bounded by two barriers induced by the mechanical deformation [99]. Coulomb blockade has been observed for this device at room temperature, with an addition energy equal to 120 meV, much higher than the thermal energy.

Another example of an SET operating at room temperature has been made using a mask technique [100]. An island about 10 nm across was fabricated in a little bundle (or rope) of nanotubes by local chemical modification. In this system oscillations were observed in the conductance as a function of the gate voltage and peaks were observed in the differential conductance as a function of the drain–source bias right up to room temperature. Note that, using a similar mask technique to carry out selective chemical doping, a nanoscale device with negative differential resistance (a  $p$ – $n$  junction) and able to operate at room temperature was fabricated on a single nanotube [101]. More recently, a  $p$ – $n$ – $p$  junction-type device was created by the same group.

The realisation of this set of components attests to the remarkable vitality of this field. Other types of component such as gas sensors or detectors of enzyme activity have also been made. The reader is referred to [102, 103] for more detail.

## 13.5 From Components to Circuits

### 13.5.1 Fabrication Techniques

The development of the first molecular components raised the problem of connecting together such components to build circuits. Today, it is too soon to seriously consider building 3D circuits. However, fabrication techniques using controlled positioning of nano-objects on a surface are moving forward rapidly. These techniques, often associated with the idea of self-assembly, may well play an important role in the future. We begin with a brief discussion of molecules and then turn to carbon nanotubes.

#### Molecular Assembly Techniques

As indicated in Sect.13.3, molecular components are based on electrode–molecule–electrode junctions. Molecules can be assembled on electrodes by the following techniques:

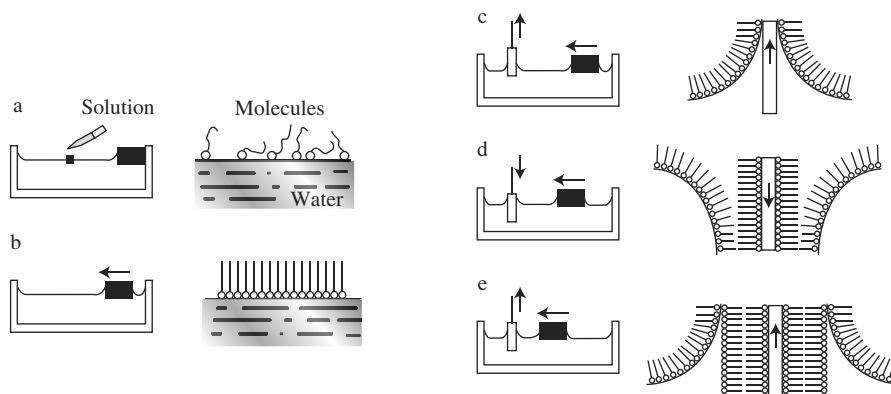
- vacuum sublimation,
- Langmuir–Blodgett (LB) technique,
- reactive self-assembly.

##### *Langmuir–Blodgett Technique*

The basic idea of the LB technique invented by I. Langmuir and K. Blodgett in the 1930s will be outlined below. The reader should refer to the book by Roberts et al. [104] for further details concerning this supramolecular engineering technique which received much attention during the 1980s in particular.

Figure 13.38 is a schematic representation of the underlying idea. Amphiphilic molecules containing a hydrophilic polar head and a hydrophobic tail are dissolved in a volatile solvent and spread across the surface of water in an LB tank. The solvent is then evaporated and the molecules are compressed by a mobile barrier which gradually reduces the area available to them. Once this compression phase has been achieved, a 2D monomolecular film (usually solid) is formed at the air–water interface. This film can be transferred, monomolecular layer by monomolecular layer, onto a solid substrate by passing the latter through the air–water interface. One layer is deposited each time the interface is crossed, in such a way that its hydrophobic side (resp. hydrophilic side) is in contact with the substrate or the substrate covered with layers exposing a hydrophobic surface (resp. hydrophilic surface).

This is an extremely versatile technique, especially since it was extended to hydrophobic molecules in the 1990s, and it has the advantage that it does not require the formation of a chemical bond between the molecules and the substrate, in contrast to the self-assembly technique described below. However, the lateral dimensions of the deposited film must be defined a posteriori and furthermore, LB films are not always completely free of defects (holes) which may lead to electrical defects in electrode–molecule–electrode structures.

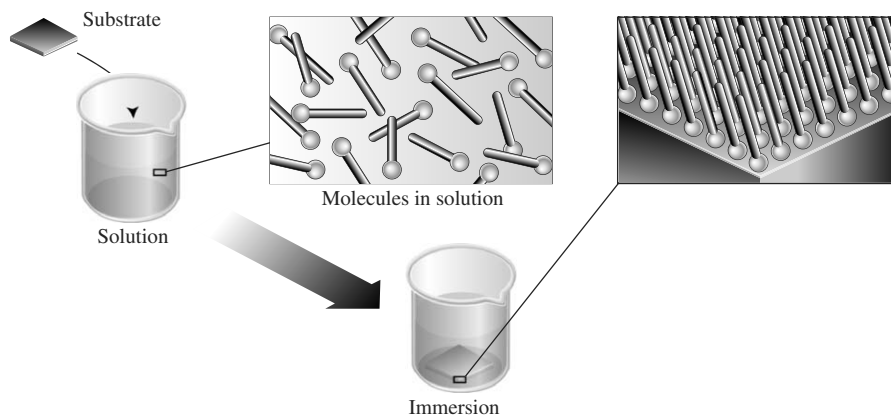


**Fig. 13.38.** Formation of a monomolecular film. (a) A solution of amphiphilic molecules is spread on water and then the solvent is evaporated to leave the molecules fixed on the water surface. (b) The film is compressed isothermally using a mobile barrier. (c)–(e) The film is transferred onto a solid substrate (hydrophilic in the example) which passes through the monomolecular film at the air–water interface: (c) first layer, (d) second layer, (e) third layer

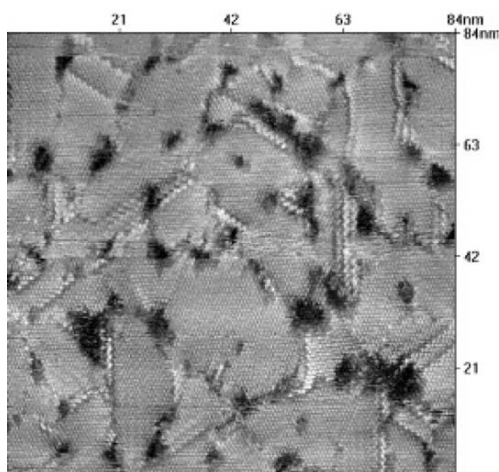
#### *Self-Assembled Monomolecular Layers*

The idea underlying the formation of self-assembled monomolecular (SAM) layers is illustrated in Fig. 13.39. A substrate is immersed in a solution of molecules which carry a reactive group with respect to the substrate. The molecules adsorb chemically onto the substrate surface. Depending on the commensurability between the lattice structures of the substrate surface and the crystallised molecules, the mobility of the molecules adsorbed onto the surface, and the nature of the interactions between these molecules, a monomolecular layer with varying degrees of density and order will result. For example, with self-assembled layers of alkylthiol (for an aliphatic chain of length at least 10 carbon atoms) on a gold substrate, dense domains with crystalline order are often observed over areas as large as several hundred square nanometers, as can be seen from the STM image in Fig. 13.40. Table 13.2 shows some of the many pairs of reactive groups and surfaces suitable for elaborating SAMs.

SAMs are very easy to localise spatially (in the lateral sense). One approach is by chemical or mechanical structuring of the substrate in such a way as to limit the regions where deposition occurs. Another is to bring in reactive molecules in a local manner on a stamp (microcontact printing, see Chap. 6), a technique developed by G.M. Whitesides at Harvard university (USA) [105], or on an AFM tip (dip-pen lithography, see Chap. 6), a technique developed by C. Mirkin at Northwestern [106]. In this approach the stamp or tip is said to be inked with the molecules, by analogy with conventional printing. It has also been shown that these layers can be lithographed a posteriori by an AFM tip, STM tip, or electron beam.



**Fig. 13.39.** Formation of self-assembled monomolecular layers by immersion of a substrate in a solution of molecules that are reactive with respect to this substrate



**Fig. 13.40.** STM image of a self-assembled layer of dodecanethiol on gold (111). Each *white dot* corresponds to a molecule. *Dark regions* correspond to holes in the gold surface

Note also that it is possible to form multilayers by using bifunctional molecules. The reader is referred to the review article on self-assembled layers by A. Ulmann [107].

### Assembly Techniques for Carbon Nanotubes and Nanowires

As discussed in Sect. 13.4, many devices have been made using carbon nanotubes, and some with quite remarkable performance. These are generally made by one of the following methods:

**Table 13.2.** Examples pairs of reactive groups and substrates that can be used to form SAMs. R denotes an alkyl or aryl radical carrying one or more heteroatoms, for example

Substrate	Molecule	Monolayer
Au	RSH	RS-Au
Au	RSR'	RS-Au
Au	RSO <sub>2</sub> H	RSO <sub>2</sub> -Au
Au	R <sub>3</sub> P	R <sub>3</sub> P-Au
Ag	RSH	RS-Ag
Cu	RSH	RS-Cu
Pd	RSH	RS-Pd
Pt	RNC	RNC-Pt
GaAs	RSH	RS-GaAs
InP	RSH	RS-InP
SiO <sub>2</sub>	RSiCl <sub>3</sub> , RSi(OR') <sub>3</sub>	Siloxane
Si/SiH	(RCOO) <sub>2</sub>	R-Si
Si/SiH	RCH=CH <sub>2</sub>	RCH <sub>2</sub> CH <sub>2</sub> Si
Si/SiCl	RLi, RMgX	R-Si
Metal oxide	RCOOH, RSiCl <sub>3</sub>	RCOO...MO <sub>n</sub>
Metal oxide	RCONHOH	RCONHOH...MO <sub>n</sub>
ZrO <sub>2</sub>	RPO <sub>3</sub> H <sub>2</sub>	RPO}...Zr
ITO	RPO <sub>3</sub> H <sub>2</sub>	RPO}...M

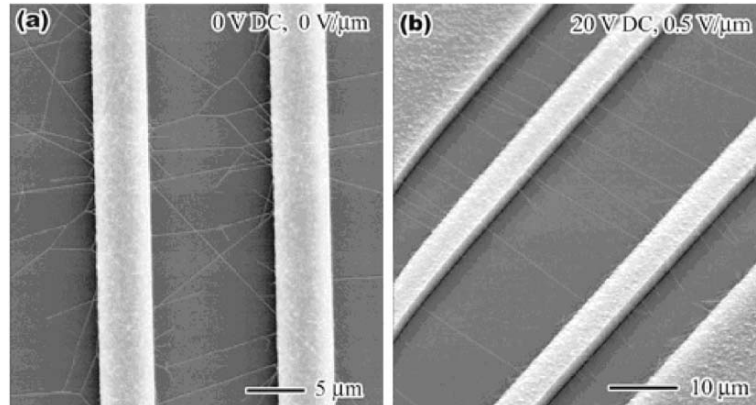
- Random deposition of nanotubes on an array of metal electrodes already produced on the surface by lithography.
- Positioning the nanotubes on or between the metal electrodes by means of an AFM tip.
- Random deposition of nanotubes directly on the silicon substrate. In this case, each nanotube must first be located on the surface before it can be connected to electrodes made by electron lithography.

To fabricate devices in a systematic and reproducible way and move from the level of the individual component to full-scale circuits, one therefore requires a controlled way of assembling the nanotubes on a surface. Various solutions have been envisaged. It should be noted that most of these solutions apply equally well to the field of metallic or semiconducting nanowires. For more information concerning this very promising area of development, the reader is referred to the review article by C.M. Lieber [108].

#### *Localised Growth of Nanotubes by CVD*

Chemical vapor deposition (CVD) is an in situ technique for synthesising nanotubes and their connections to metal measuring electrodes ( $T_{\text{synthesis}} \approx 100^\circ\text{C}$ ) in a single step (see Chap. 8). In addition, it has been shown that the direction of growth of the nanotubes can be controlled by applying an





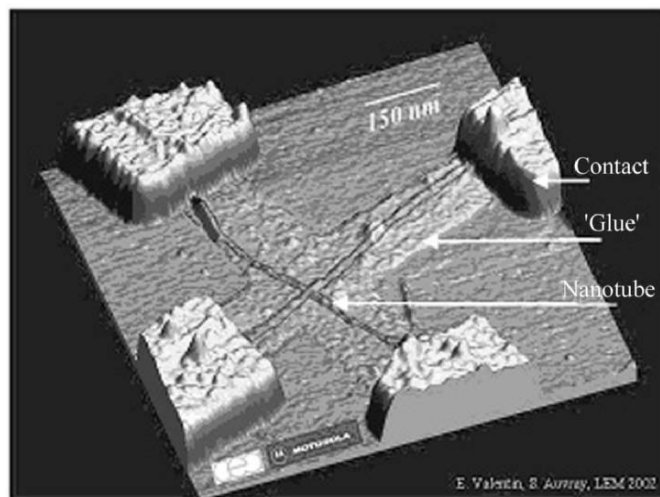
**Fig. 13.41.** MEB images [109] of nanotubes synthesised by CVD and suspended between two electrodes. (a) Random growth of nanotubes. (b) Nanotube growth in a predetermined direction by applying a constant electric field of  $0.5 \text{ V}/\mu\text{m}$

electric field between the dots of metal catalysts during the synthesis (see Fig. 13.41) [109]. This method provides a quick way of fabricating many devices. Moreover, it avoids all the purification and sonication stages which degrade the surface of the nanotubes before they go into use. But although the CVD approach is a good solution for fabricating high performance CNT-FET devices, it does involve a certain number of drawbacks. For example, the diameter and chirality of the nanotubes cannot be controlled during synthesis. Furthermore, The positioning of the catalyst must be controlled, even at the temperature of synthesis, which seems to seriously limit development prospects.

#### *Controlled Deposition of Nanotubes in Solution*

The carbon nanotubes used in this case are synthesised by high temperature methods ( $T_{\text{synthesis}} \approx 1200^\circ\text{C}$ ) such as laser ablation or electric arc. With these techniques, the nanotube diameter can be precisely controlled during synthesis (in the range 1–1.3 nm), thereby reducing the energy distribution of the band gap of semiconducting nanotubes (in the range 600–700 meV). Once they have been purified, the nanotubes are used in the form of dispersions in solvents. These comprise a statistical mixture of 1/3 metallic nanotubes and 2/3 semiconducting nanotubes. Very recent studies have revealed encouraging prospects for chemically separating metallic and semiconducting nanotubes in solution [110, 111]. If these results are confirmed, techniques available for controlling the adsorption of pre-sorted nanotubes (into semiconducting or metallic) onto surfaces may open the way to large-scale and cheap production of devices.

The idea of localised deposition is based on the local functionalisation of predetermined regions of the substrate by means of a self-assembled



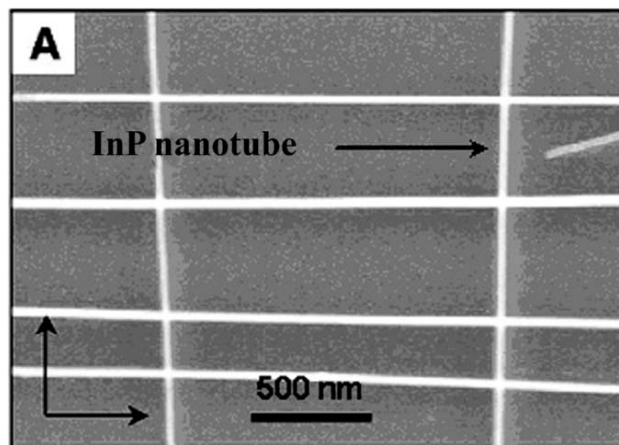
**Fig. 13.42.** Carbon nanotubes deposited selectively from a solution of NMP on a silicon substrate that has been locally functionalised by a monomolecular layer serving as a kind of glue. Metal contacts are fabricated subsequently by lithography [114]

molecular monolayer. Depending on the properties of the monolayer (hydrophilic, hydrophobic, positively or negatively charged), the nature of the nanotube/surface interaction will be modified, thereby inducing a preferential and controlled adsorption of nanotubes onto selected regions [112–114]. Figure 13.42 shows an example of crossed nanotubes made by this technique. A monomolecular layer of aminopropyl triethoxysilane is used. The carbon nanotubes dispersed in a solution of *N*-methyl-2-pyrrolidone (NMP) have a much greater affinity with this layer than they have with the uncoated silicon substrate.

Two further techniques have been developed to control the positioning of nanotubes or nanowires on a surface.

- Gerdes et al. [115] combined the surface functionalisation technique with another known as combing. When the solution dewets the silica surface, the nanotubes at the contact interface<sup>8</sup> (triple line) deform the dewetting front. The capillary force induced by this elastic deformation causes the alignment and combing of the nanotubes on the surface. More recently, this technique has been adapted to fabricate nanostructures with InP, GaP or Si nanowires or carbon nanotubes by microfluidics [116]. The idea behind this technique is to circulate the nanowire suspension through a network of microchannels formed by pressing an etched poly(dimethylsiloxane)

<sup>8</sup> The contact interface is the interface between the three phases: liquid (the nanotube suspension), solid (the substrate), and gaseous.



**Fig. 13.43.** MEB image [116] of a sequential deposition of InP nanowires oriented along two different directions by microfluidics. The horizontal nanowires were deposited first and then the substrate was rotated through  $90^\circ$  in order to deposit the vertical nanowires

(PDMS) stamp onto a flat silica substrate. The nanowires are adsorbed onto the surface parallel to the direction of flow under the effect of capillary forces. In certain cases, the nanowire/surface interaction is strong enough to be able to carry out a sequence of flows in different directions without the positions of previously adsorbed nanowires being affected (see Fig. 13.43).

- More recently, Diehl et al. [117] have used a.c. dielectrophoresis to fabricate complex nanostructures such as rope networks. (Nanotubes in suspension in a solvent are aligned under the effect of an alternating electric field.) An interesting result with this method is that the spacing between nanotube ropes can be adjusted and depends on their linear charge density. This effect arises from the repulsive interaction between the nanotube ropes, positively charged when dispersed in ortho-dichlorobenzene (ODCB). Once they have been deposited on the silica surface, the nanotubes modify the electric field around them. In this way, the following nanotubes are adsorbed in regions where the Coulomb repulsion is minimal. Despite promising prospects, it seems that this technique is restricted to nanotube ropes, which rather limits applications at the present time.

The projects discussed above give a rather global view of the way research has been going since 1998 with regard to the problem of controlled positioning of nanotubes on surfaces. They show that several approaches centering on the idea of self-assembly have been investigated and show promise for the step from individual components up to full-scale circuits.

### 13.5.2 Circuit Architecture

As we have seen, several types of basic molecular component have recently been developed. The next problem is to put them to use in circuits. There are two suitable architectures: a hybrid architecture, the main subject here, and an integrated architecture.

The hybrid architecture is conceptually simpler because it is closer to what is done today with the CMOS. The basic components are interconnected by means of metal wires. This applies, for example, to the case of nanotube transistors which have been used to make logic gates and small circuits, but also diode–transistor-type logic circuits, or circuits using single-electron components. These types of architecture can be tested in research establishments today. In this category, we may also classify programmable architectures based on networks of crossed nanotubes, with a switch placed at the crossing point. The big advantage with a programmable architecture is to be able to tolerate imperfections introduced during fabrication. The recent demonstration of programmable switches using layers of rotaxanes, oxidised or otherwise by prior application of a potential difference [118], and also nanotube switches [119], are in this respect important steps towards producing, e.g., prototype 64-bit memories (on an area of order  $1\ \mu\text{m}^2$ ).

This type of technology may have great advantages as far as production costs are concerned because it uses self-assembly and parallel fabrication techniques. However, it does also have two big drawbacks when it comes to large-scale use with a high level of miniaturisation, unless there is a paradigm change in the approach to computation (the ultimate CMOS will eventually encounter similar problems):

- Below about 10-nm spacing between the basic components, we may expect a loss of individual identity and interference effects between the various paths available for going from one point to another in the circuit, especially if the circuit is highly regular as in the case of regular criss-cross networks.
- Even assuming perfect switching, i.e., each operation requires an energy  $kT \ln 2$ , a circuit assembling components operating at 10 GHz with a density of a few times  $10^{12}\ \text{cm}^{-2}$  will require a power input of the order of several tens of  $\text{W}/\text{cm}^2$ , which is the limit of the power that can be dissipated technologically.

Integrated molecular electronics (monomolecular electronics), wherein several functions required for a calculation are integrated within a single molecule so that only the result is transmitted to the macroscopic world, may in principle be able to overcome some of these difficulties. The technical difficulty involved in implementing this type of architecture nevertheless remains formidable. The first theoretical steps have been taken, showing that the association rules for molecular building blocks likely to be able to carry out calculation functions when associated with a macromolecule are far from simple. For example, Kirchoff's laws do not apply and the molecule must be thought of as a whole

entity [29]. One thing is clear from these studies: the concept of a logic circuit within a molecule comes up against significant problems, such as low fan-out or a difficulty in defining well separated logic levels [120,121]. It is much more interesting to consider non-standard approaches to computation.

Cellular automata fall into this category. The idea here is to use cells of four interacting quantum dots in a square arrangement and charged with two electrons. The Coulomb repulsion forces the electrons to position themselves as far apart as possible, which gives two stable positions, namely, the two diagonals. Calculations are then carried out using the interactions induced on a cell by its neighbours. This idea has been studied theoretically in some detail and has the advantage that it only requires a very small number of connections with the outside world. However, it is a different matter to actually make this device, with apparently unrealistic constraints being imposed on the reproducibility of the quantum dots. An attempt has nevertheless been made using DNA tiles, taking advantage of the good self-assembly properties inherent in biological objects.

Finally, it seems that integrated molecular electronics can also be coupled in a natural way with quantum calculation and the first theoretical studies are under way [122].

### 13.6 Conclusion

Studies carried out today in molecular electronics, based on recent spectacular breakthroughs with the development of the first functional components and circuits, can be viewed in the following way.

From a fundamental standpoint, the aim is primarily to achieve a new transport regime by connecting nano-objects directly to conducting electrodes. This method would make it possible, in the case of a single molecule, to carry out spectroscopy on its electronic states in a direct way in the various regimes of coupling with the electrodes. By doing this, one may hope to understand and control the relationship between molecular structure and transport properties and hence, in the long run, to exploit this relationship. The first steps have already been taken in this direction, refining our understanding of this relationship.

From the applications standpoint, molecular electronics could bring to our present devices all the advantages of nano-objects that can be attributed a function chemically, and it may in the long term provide a complement to today's silicon-based microelectronics. It may allow us to go beyond the 10 nm barrier which stands in the way of the quest to reduce component size. But apart from miniaturisation, its most obvious feature, molecular electronics can also provide alternative solutions when it comes to reducing production costs, because it goes hand-in-hand with self-assembly and bottom-up techniques. And it can provide another way of dealing with complexity, by developing

novel computation architectures, and a reduction in the energy cost of computation.

Finally, let us note that, although molecular electronics naturally involves nanometric dimensions due to the kind of object it exploits, it could not be fully developed without gaining control over much smaller dimensions, of the order of the accuracy of chemical bonds. It thus seems probable that it will give rise by continuity to the field of picotechnology.

## References

1. Aviram, A., and Ratner, M.A.: Chem. Phys. Lett. **29**, 277 (1974)
2. Carter, F.L.: In *2nd Intl Symp. Molecular Electronic Devices* (M. Dekker, New York, 1982) p. 149
3. Martin, A.S., Sambles, J.R., and Ashwell, G.J.: Phys. Rev. Lett. **70**, 218 (1993)
4. Metzger, R.M., et al.: J. Am. Chem. Soc. **119**, 10455 (1997)
5. Taube, H.: Ann. N.Y. Acad. Sci. **313**, 481 (1978)
6. Effenberger, F., et al.: Angew. Chem. Int. Ed. Engl. **27**, 281 (1988)
7. Joachim, C., and Launay, J.P.: J. Molec. Electronics **6**, 37 (1990)
8. Launay, J.P., et al.: Inorg. Chem. **30**, 1033 (1991)
9. Joachim, C., and Launay J.P.: Chem. Phys. **109**, 93 (1986)
10. Hauser, A.: Chem. Phys. Lett. **124**, 543 (1986)
11. Joachim, C., and Gimzewski, J.K.: Europhysics Letters **30**, 409 (1995)
12. Bourgoin, J.-P.: In *Interacting Electrons in Nanostructures*, ed. by H.S. Schoeller and R. Haug (Springer Verlag, Berlin, 2001)
13. Joachim, C., Gimzewski, J.K., and Aviram A.: Nature **408**, 541 (2000)
14. Bumm, L.A., et al.: Science **271**, 1705 (1996)
15. Kergueris, C., Bourgoin, J.P., and Palacin, S.: Nanotechnology **10**, 8 (1999)
16. Leatherman, G., et al.: J. Phys. Chem. B **103**, 4006 (1999)
17. Patrone, L., et al.: Chem. Phys. **281**, 325 (2002)
18. Langlais, V.J., et al.: Phys. Rev. Lett. **83**, 2809 (1999)
19. Joachim, C., and Gimzewski, J.K.: Europhys. Lett. **30**, 409 (1995)
20. Nitzan, A., and Ratner, M.A.: Science **300**, 1384 (2003)
21. Reed, M.A., et al.: Science **278**, 252 (1997)
22. Muller, C.J., et al.: Nanotechnology **7**, 409 (1996)
23. Kergueris, C., et al.: Phys. Rev. B **59**, 12505 (1999)
24. Reichert, J., et al.: Phys. Rev. Lett. **88**, 176804 (2002)
25. van Ruitenbeek, J.M., et al.: Rev. Sci. Instrum. **67**, 108 (1995)
26. Andres, R.P., et al.: Science **273**, 1690 (1996)
27. Bourgoin, J.P., et al.: **327-329**, 515 (1998)
28. McConnell, W.P., et al.: J. Phys. Chem. B **104**, 8925 (2000)
29. Magoga, M., and Joachim, C.: Phys. Rev. B **59**, 16011 (1999)
30. Yaliraki, S.N., and Ratner, M.A.: J. Chem. Phys. **109**, 5036 (1998)
31. Chen, J., et al.: Chem. Phys. Lett. **313**, 741 (1999)
32. Chen, J., et al.: Science **286**, 1550 (1999)
33. Chen, J., et al.: Appl. Phys. Lett. **77**, 1224 (2000)
34. Zhou, C., et al.: Appl. Phys. Lett. **71**, 611 (1997)
35. Bezryadin, A., and Dekker, C.: J. Vac. Sci. Technol. B **15**, 793 (1997)

36. Porath, D., et al.: *Nature* **403**, 635 (2000)
37. Morpurgo, A.F., Marcus, C.M., and Robinson, D.B.: *Appl. Phys. Lett.* **74**, 2084 (1999)
38. Olofsson, L.G.M., et al.: *J. Low Temp. Phys.* **118**, 343 (2000)
39. Park, H., et al.: *Appl. Phys. Lett.* **75**, 301 (1999)
40. Park, H., et al.: *Nature* **407**, 57 (2000)
41. Lambert, M.F., et al.: *Nanotechnology* **14**, 772 (2003)
42. Porath, D., and Millo, O.: *J. Appl. Phys.* **81**, 2241 (1997)
43. Porath, D., et al.: *Phys. Rev. B* **56**, 9829 (1997)
44. Joachim, C., et al.: *Phys. Rev. Lett.* **74**, 2102 (1995)
45. Cramer, C.J.: *Essentials of Computational Chemistry: Theories and Models* (Wiley, New York, 2002)
46. Patrone, L., et al.: *Phys. Rev. Lett.* **91**, 096802 (2003)
47. Mujica, V., Roitberg, A.E., and Ratner, M.: *J. Chem. Phys.* **112**, 6834 (2000)
48. Lenfant, S., et al.: *Nano. Lett.* **3**, 741 (2003)
49. Damle, P., Ghosh, A.W., and Datta, S.: *Chem. Phys.* **281**, 171 (2002)
50. Desjonqueres, M.C., and Spanjaard, D.: *Concepts in Surface Physics* (Springer-Verlag, Berlin, New York, Heidelberg, 1996)
51. Bonet, E., Deshmukh, M.M., and Ralph, D.C.: *Phys. Rev. B* **65** (2002)
52. Hettler, M.H., et al.: *Phys. Rev. Lett.* **90**, 076805 (2003)
53. Adams, D.M., et al.: *J. Phys. Chem. A* **107**, 6522 (2003)
54. Salomon, A., et al.: *Adv. Mater.* (Weinheim, Germany) in press (2003)
55. Kushmerick, J.G., et al.: *J. Am. Chem. Soc.* **124**, 10654 (2002)
56. Yaliraki, S.N., and Ratner, M.A.: *Abstr. Pap. Am. Chem. Soc.* **218**, 155 (1999)
57. Di Ventra, M., and Lang, N.D.: *Phys. Rev. B* **65**, 045402 (2002)
58. Taylor, J., Brandbyge, M., and Stokbro, K.: *Phys. Rev. Lett.* **89**, 057904 (2002)
59. Martin, A.S., Sables, J.R., and Ashwell, G.J.: *Phys. Rev. Lett.* **70**, 218 (1993)
60. Metzger, R.M., et al.: *J. Am. Chem. Soc.* **119**, 10455 (1997)
61. Metzger, R.M., Xu, T., and Peterson, I.R.: *J. Phys. Chem. B* **105**, 7280 (2001)
62. Vuillaume, D., Chen, B., and Metzger, R.M.: *Langmuir* **15**, 4011 (1999)
63. Xu, T., et al.: *Angew. Chem. Int. Ed. Engl.* **40**, 1749 (2001)
64. Krzeminski, C., et al.: *Phys. Rev. B* **64**, 085405 (2001)
65. Kornilovitch, P.E., Bratkovsky, A.M., and Williams, R.S.: *Phys. Rev. B* **66**, 165436 (2002)
66. Collet, J., et al.: *Appl. Phys. Lett.* **76**, 1339 (2000)
67. Vuillaume, D.: *J. Nanosci. Nanotech.* **2**, 267 (2002)
68. Krzeminski, C.: *Doctoral thesis, University of Lille, France* (2001)
69. Joachim, C., and Gimzewski, J.K.: *Chem. Phys. Lett.* **265**, 353 (1997)
70. Schill, G.: *Catenanes, Rotaxanes and Knots* (Academic Press, New York, 1971)
71. Sauvage, J.P.: *Accounts Chem. Res.* **23**, 319 (1990)
72. Bissel, R.A., et al.: *Nature* **369**, 133 (1994)
73. Lehn, J.M.: *La chimie supramoléculaire, concepts et perspectives* (Deboeck University, Paris, 1997)
74. Collier, C.P., et al.: *Science* **289**, 1172 (2000)
75. Collier, C.P., et al.: *Science* **285**, 391 (1999)
76. Pease, A.R., et al.: *Accounts Chem. Res.* **34**, 433 (2001)
77. Chen, Y., et al.: *Nanotechnology* **14**, 462 (2003)
78. Chen, Y., et al.: *Appl. Phys. Lett.* **82**, 1610 (2003)
79. Di Ventra, M., Pantelides, S.T. and Lang, N.D.: *Appl. Phys. Lett.* **76**, 3448 (2000)

80. Tans, S.J., Verschueren, A.R.M., and Dekker, C.: *Nature* **393**, 49 (1998)
81. Martel, R., et al.: *Appl. Phys. Lett.* **73**, 2447 (1998)
82. Bachtold, A., et al.: *Science* **294**, 1317 (2001)
83. Bockrath, M., et al.: *Phys. Rev. B* **61**, 10606 (2000)
84. Derycke, V., et al.: *Nano. Lett.* **1**, 453 (2001)
85. Wind, S.J., et al.: *Appl. Phys. Lett.* **80**, 3817 (2002)
86. Rosenblatt, S., et al.: *Nano. Lett.* **2**, 869 (2002)
87. Javey, A., et al.: *Nature Materials* **1**, 241 (2002)
88. McEuen, P.L., Fuhrer, M.S., and Park, H.K.: *IEEE Trans. Nanotech.* **1**, 78 (2002)
89. Sze, S.: *Physics of Semiconductor Devices* (Wiley, New York, 1981)
90. Heinze, S., et al.: *Phys. Rev. Lett.* **89**, 106801 (2002)
91. Leonard, F., and Tersoff, J.: *Phys. Rev. Lett.* **84**, 4693 (2000)
92. Odintsov, A.A.: *Phys. Rev. Lett.* **85**, 150 (2000)
93. Appenzeller, J., et al.: *Phys. Rev. Lett.* **89**, 126801 (2002)
94. Appenzeller, J., et al.: *IEEE Trans. Nanotech.* **1**, 184 (2002)
95. Hoenlein, W., et al.: *Mat. Res. Soc. Symp. Proc.* **772**, M4.5.1 (2003)
96. Grabert, H., and Devoret, M.H.: *Single Charge Tunneling* (Plenum Press, New York, 1992)
97. Tans, S.J., et al.: *Nature* **386**, 474 (1997)
98. Postma, H.W.C., Yao, Z., and Dekker, C.: *J. Low Temp. Phys.* **118**, 495 (2000)
99. Postma, H.W.C., et al.: *Science* **293**, 76 (2001)
100. Cui, J.B., Burghard, M., and Kern, K.: *Nano. Lett.* **2**, 117 (2002)
101. Zhou, C., et al.: *Science* **290**, 1552 (2000)
102. Dai, H.: *Accounts Chem. Res.* **35**, 1035 (2002)
103. Avouris, P.: *Accounts Chem. Res.* **35**, 1026 (2002)
104. Roberts, G.L.: *Langmuir–Blodgett Films* (Plenum Press, New York, 1990)
105. Xia, Y., et al.: *Chem. Rev.* **99**, 1823 (1999)
106. Piner, R.D., et al.: *Science* **283**, 661 (1999)
107. Ulman, A.: *Chem. Rev.* **96** (4), 1533 (1996)
108. Lieber, C.M.: *MRS Bull.* **28**, 486 (2003)
109. Zhang, Y.G., et al.: *Appl. Phys. Lett.* **79**, 3155 (2001)
110. Chattopadhyay, D., Galeska, L., and Papadimitrakopoulos, F.: *J. Am. Chem. Soc.* **125**, 3370 (2003)
111. Krupke, R., et al.: *Science* **301**, 344 (2003)
112. Liu, J., et al.: *Chem. Phys. Lett.* **303**, 125 (1999)
113. Choi, K.H., et al.: *Surf. Sci.* **462**, 195 (2000)
114. Valentin, E., et al.: *Microelectron. Eng.* **61–2**, 491 (2002)
115. Gerdes, S., et al.: *Europhys. Lett.* **48**, 292 (1999)
116. Huang, Y., et al.: *Science* **291**, 630 (2001)
117. Diehl, M.R., et al.: *Angew. Chem. Intl. Ed.* **41**, 353 (2002)
118. Collier, C.P., et al.: *J. Am. Chem. Soc.* **123**, 12632 (2001)
119. Rueckes, T., et al.: *Science* **289**, 94 (2000)
120. Ami, S., Hliwa, M., and Joachim, C.: *Chem. Phys. Lett.* **367**, 662 (2003)
121. Ami, S., Hliwa, M., and Joachim, C.: *Nanotechnology* **14**, 283 (2003)
122. Joachim, C.: *Nanotechnology* **13**, R1 (2002)



---

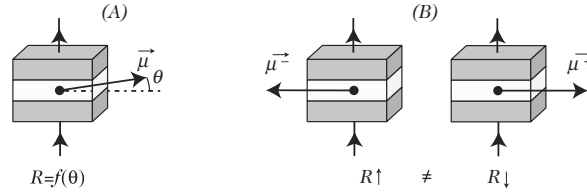
## Nanomagnetism and Spin Electronics

C. Chappert and A. Barthélémy

In the year 2000, as many hard disks were sold for computing as televisions. And new applications for the general public, such as decoder-recorders, video players, and juke boxes, have just appeared on the scene. This technology, now so widespread, represents one of the first examples of nanotechnology, where the need for mass production must be reconciled with the fundamental limits of nanomagnetism, i.e., fundamental magnetic processes on the nanometric scale. Indeed, the successful storage of information now rests upon the thermal stability of magnetic grains containing fewer than 100 000 atoms. And it is by controlling the properties of these grains that it has been possible to increase the recording surface density by a factor of  $10^7$  since the first hard disks sold by IBM in 1957.

This exponential increase in performance has been carried along by continual improvements associated with a long series of technological breakthroughs. One of the most fruitful was the discovery in 1988 of giant magnetoresistance in multilayers, made independently by A. Fert in France and P. Grünberg in Germany [1, 2]. This phenomenon can be used to relate the magnetic configuration and electrical resistance with a new level of sensitivity, opening the way to a significant miniaturisation of certain devices. Used since 1997 in read heads for hard disks (see Fig. 14.1A), this discovery also lies at the heart of the new magnetic random access memory (MRAM) (see Fig. 14.1B), whilst yet other devices are in the pipeline that will push magnetism still further into the world of electronics. And so we come to the new area of spin electronics, or spintronics, the subject of this chapter.

Recording applications are treated in detail in Chap. 15. The aim of this chapter is not therefore to provide a complete presentation of magnetism and the reader is referred to the literature [3]. Instead, we shall describe the basic phenomena underlying the behaviour of spintronic devices. In one sentence, this may be summed up as follows: on the scale of a few nanometers, magnetic objects come into the range of the fundamental lengths of magnetism and electron transport, which gives rise to novel effects that are heavily dependent on size. Giant magnetoresistance is an example. It is also an area where



**Fig. 14.1.** A magneto-electronic device can be defined quite generally as a multilayer device whose resistance  $R$  varies with the orientation of the magnetic moment  $\mu$  of one of the layers. **(A)** To measure a magnetic field (sensor function), this layer is left relatively free to rotate under the effect of the field. **(B)** To store binary information in MRAM magnetic memories, the magnetic moment must be restricted to two allowed antiparallel orientations, representing 0 and 1

surface and interface effects play a crucial role. Finally, the dynamics of these processes, an important parameter for applications, also brings in characteristic times of the order of the nanosecond. We shall see how these fundamental quantities arise from the microscopic processes of magnetism and electrical conduction in ferromagnetic metals, and we shall illustrate the main effects by examples drawn from recent progress in research.

## 14.1 Nanomagnetism

### 14.1.1 Vacuum Magnetostatics

As for electrostatics, the theory of magnetism can be built upon the study of magnetic forces. In vacuum, these forces arise purely from the electric current distributions, and the principle of superposition, i.e., the fact that the effects of several sources are additive, then allows one to describe them by introducing the magnetic field  $\mathbf{H}$  (in units of A/m) and magnetic induction  $\mathbf{B}$  (in units of T). These are proportional, i.e.,  $\mathbf{B} = \mu_0 \mathbf{H}$ , where  $\mu_0 = 4\pi \times 10^{-7} \text{ H m}^{-1}$ .

The most important idea is then the magnetic dipole moment  $\mu$  associated with an infinitesimal current loop with directed surface element  $\mathbf{S}$  and carrying current  $I$ :

$$\mu = IS .$$

This idea of the magnetic dipole extends to all permanent current distributions, when we study effects at distances that are much greater than the dimensions of the current distribution.

A magnetic dipole moment creates a magnetic induction  $\mathbf{b}$  in space given by

$$\mathbf{b} = \frac{\mu_0}{4\pi} \left[ \frac{3}{r^5} (\boldsymbol{\mu} \cdot \mathbf{r}) \mathbf{r} - \frac{\boldsymbol{\mu}}{r^3} \right] , \tag{14.1}$$

where  $\mathbf{r}$  is the vector from the dipole to the field point at which the induction is evaluated.

The word ‘dipole’ is used by analogy with the electric dipole moment, due to the perfect similarity between (14.1) and the expression for the electric field created by an electric dipole  $\mathbf{p} = q\mathbf{l}$ , where  $\mathbf{l}$  is the vector from a charge  $-q$  to a charge  $+q$ . The analogy is obtained by replacing  $\mathbf{m}$  by  $\mathbf{p}$  and  $\mu_0$  by  $1/\varepsilon_0$ . This is an important analogy, because with the right transformations, it can be used to calculate the magnetic inductions using the simpler methods of electrostatics, and in particular the idea of a scalar potential. However, this is only an analogy, and there remains a fundamental difference in the context of relativistic quantum mechanics: the magnetic dipole moment changes sign under time reversal, in contrast to the electric dipole.

Let us suppose now that this dipole  $\boldsymbol{\mu}$  is situated in a uniform magnetic induction  $\mathbf{B}$ . The couple exerted by  $\mathbf{B}$  on  $\boldsymbol{\mu}$  arises directly from the Laplace force (action of  $\mathbf{B}$  on a circuit element with a current in it):

$$\boldsymbol{\Gamma} = \boldsymbol{\mu} \times \mathbf{B} . \quad (14.2)$$

The expression for the corresponding energy is generally used:

$$E_B = -\mu_0(\boldsymbol{\mu} \cdot \mathbf{H}) . \quad (14.3)$$

This expression is valid in the very general case of a ‘rigid’ dipole, whose moment  $\boldsymbol{\mu}$  does not depend on  $\mathbf{H}$ . The minimum energy is thus reached when  $\boldsymbol{\mu}$  is aligned with  $\mathbf{H}$ .

If another magnetic dipole is located at  $\mathbf{r}$ , the interaction energy between the two dipoles, denoted  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\mu}_2$ , can be found immediately from (14.1) and (14.3):

$$E_{\text{dip}} = \frac{\mu_0}{4\pi} \left[ \frac{\boldsymbol{\mu}_1 \cdot \boldsymbol{\mu}_2}{r^3} - 3 \frac{(\boldsymbol{\mu}_1 \cdot \mathbf{r})(\boldsymbol{\mu}_2 \cdot \mathbf{r})}{r^5} \right] . \quad (14.4)$$

The dipole interaction is not the strongest in terms of magnetic energy. However, it has a considerable importance owing to its anisotropy, i.e., the fact that it depends on the relative orientation of  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\mu}_2$  and their orientations relative to  $\mathbf{r}$ , and its long range, which means that all the moments of a material contribute to the energy at a given point.

### 14.1.2 Magnetism in Matter: Fundamental Relations

Matter is also a place of charged particle motion, e.g., electrons around atomic nuclei, where magnetic effects are governed by the fundamental relations of quantum mechanics. A knowledge of these fundamental rules of magnetism in matter is essential in order to understand effects observed on the nanoscale. We shall outline them here, but without going into a detailed theoretical discussion such as can be found in more specialised textbooks [4]. We shall limit ourselves to magnetism of electronic origin, since this is what gives rise to the magnetic effects that are relevant in the present chapter.

### A Relativistic Quantum Effect

Non-relativistic quantum theory treats the electron as a material point with three degrees of freedom associated with its spatial coordinates  $x, y, z$ . In this framework, an electronic quantum state is characterised by a spatial wave function  $\psi(x, y, z)$  which is a solution of the Schrödinger equation. The probability of the electron being at the point  $x, y, z$  is then equal to  $|\psi|^2$ . Applying the principles of relativity leads to the Dirac equation, which introduces a further quantity characterising the electron, namely its spin. This quantity is an intrinsic property of various particles, like their rest mass. The idea of spin was originally introduced by Uhlenbeck and Goudsmit (1925), who proposed that the electron might rotate about its own axis, hence conferring a spin angular momentum upon it. In quantum theory, the spin angular momentum  $\mathbf{s}$  of an electron projected along any predetermined axis assumes the values  $\pm 1/2$ . The quantum state of an electron is then specified by its spatial and spin states.

#### *Magnetic Moment of an Electron*

Consequently, the magnetic moment associated with an electron rotating about an atomic nucleus has two sources. Associated with the orbital angular momentum  $\mathbf{l}$  of the electron in its spatial state, there is an orbital magnetic moment  $\boldsymbol{\mu}_l$  given by

$$\boldsymbol{\mu}_l = -\frac{\mu_B}{\hbar} \mathbf{l}, \quad (14.5)$$

where  $\mu_B = 0.9273 \times 10^{-23} \text{ A m}^2$  is the Bohr magneton. This corresponds exactly to the classical definition, where the electron rotating about the nucleus is treated as an infinitesimal current loop with quantised angular momentum  $\mathbf{l}$ .

In a like manner, associated with the spin angular momentum  $\mathbf{s}$ , there is a spin magnetic moment  $\boldsymbol{\mu}_S$  given by

$$\boldsymbol{\mu}_S = -g \frac{\mu_B}{\hbar} \mathbf{s}, \quad (14.6)$$

where  $g$  is the gyromagnetic ratio, very close to 2. The projection of  $\boldsymbol{\mu}_S$  onto the spin angular momentum of the electron is almost equal to  $\pm \mu_B$ .

#### *Spin-Orbit Interaction*

Another consequence of the principles of relativity is the existence of a spin-orbit interaction which couples the orbital and spin angular momenta. The corresponding Hamiltonian is

$$H_{\text{SO}} = -\xi \mathbf{l} \cdot \mathbf{s}, \quad (14.7)$$

where  $\xi$  is the coupling constant. The classical analogy allows some understanding of this effect. In a frame of reference moving with the electron, the motion of the nucleus is equivalent to a current loop creating a magnetic field at the electron which interacts with its spin magnetic moment. This classical image predicts that the spin-orbit interaction will increase with the charge  $Z$  on the nucleus, i.e., it will be greater for heavy atoms. Even though the actual dependence is much more complex and involves other effects, this general tendency is rather well borne out. For example, the spin-orbit coupling constant for platinum (Pt,  $Z = 78$ ) is 6 times greater than that for cobalt (Co,  $Z = 27$ ) [5].

The spin-orbit interaction is fundamental because it underlies many of the effects we shall describe in this chapter, which we shall refer to under the heading of magnetocrystalline anisotropy: in a material, the total magnetic energy depends on the orientation of the magnetic moment relative to the axes of symmetry of the crystal lattice. Note that the spin-orbit interaction, as well as the magnetic anisotropy it causes, are local quantities, likely to vary rapidly from one atomic site to another.

#### *Exchange Interaction*

Quantum mechanics lies at the origin of another important magnetic effect. The indistinguishability of electrons, associated with their half-integer spin (the electron is a fermion), leads to the Pauli exclusion principle: two electrons with the same spin cannot occupy the same spatial state. In quantum terms, the probability of their being simultaneously at the same point, averaged over the whole of space, is lower than for electrons with antiparallel spins. They thus have a lower Coulomb interaction energy (the Coulomb interaction being repulsive between charges of the same sign), which leads to an effective interaction favouring a parallel alignment of the electron spins. This is the exchange interaction. In reality, the exchange interaction can be intratomic, as described above, or interatomic through valence electrons, and it can even be transmitted through an intermediate atom. It can favour a parallel alignment of spins (ferromagnetic interaction), or an antiparallel alignment (antiferromagnetic interaction). It can even change sign between nearest and next-nearest neighbours.

In the magnetic materials with which we are concerned here, the exchange interaction is one or two orders of magnitude stronger than the other sources of magnetic energy. However, because it requires atomic orbitals to overlap, it does have a very short range, of the order of the interatomic distance. It is therefore another local quantity. Moreover, it is almost isotropic and is irrelevant to the various anisotropies.

#### *Magnetic Moment of an Isolated Atom*

An atom contains several interacting electrons. In the isolated atom, these electrons are distributed over states of increasing energy (Hund rules), taking

into account the spin-orbit interaction and respecting Pauli's principle [3, 6]. The magnetic moment  $\boldsymbol{\mu}_{\text{total}}$  of the atom is then related to the total angular momentum  $\mathbf{J} = \mathbf{L} + \mathbf{S}$ , the sum of the total orbital and spin angular momenta

$$\mathbf{L} = \sum_i \mathbf{l}_i \quad \text{and} \quad \mathbf{S} = \sum_i \mathbf{s}_i ,$$

respectively, by

$$\boldsymbol{\mu}_{\text{total}} = -g_J \frac{\mu_B}{\hbar} \mathbf{J} . \quad (14.8)$$

The Landé factor  $g_J$  of an atom is given by

$$g_J = \frac{3}{2} + \frac{S(S+1) - L(L+1)}{2J(J+1)} .$$

The projection of  $\mathbf{J}$  onto the quantisation axis can take the values  $M_J = -J, -J+1, \dots, +J$ .

Very few atoms have zero magnetic moment when isolated, because apart from a few special cases the condition  $J = 0$  can only be fulfilled if all the atomic shells are full, e.g., the noble gases.

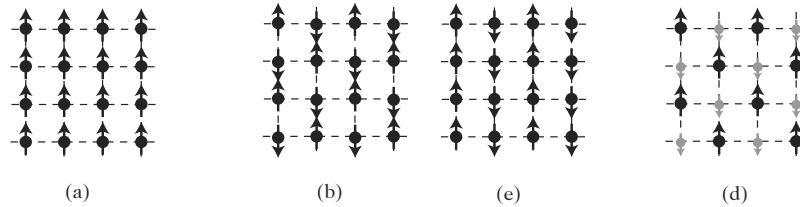
### Localised Magnetism and Itinerant Magnetism

In a material, atoms are arranged on the sites of a crystal lattice and their electrons move in the crystal field, i.e., the electric field created by all the other charges of the crystal, which therefore possesses the lattice symmetry. To a first approximation, there are two limiting cases.

#### *Localised Magnetism*

In the case of so-called localised magnetism, it is assumed that the electrons responsible for magnetic effects remain localised around their respective nuclei. In the ideal case, the crystal field can be treated as a small perturbation of the quantum state of the electron in the isolated atom (or in the isolated ion in the case of ionic crystals). The sites of the crystal lattice are then occupied by atoms carrying a magnetic moment given by (14.8), which interact mainly by the interatomic exchange interaction. Below a critical temperature, thermal agitation can no longer combat this interaction and long-range magnetic order is established (see Fig. 14.2). Depending on the type of interaction and the material (alloy, pure metal, etc.), this order may be:

- ferromagnetic, i.e., below the Curie temperature  $T_C$ , all magnetic moments are parallel,
- antiferromagnetic, i.e., below the Néel temperature  $T_N$ , two ferromagnetic sublattices with antiparallel moments balance one another to give a zero total moment,



**Fig. 14.2.** Examples of long-range magnetic order on a 2D square lattice. (a) Ferromagnetic order. All moments are aligned parallel to one another. (b) and (c) Antiferromagnetic order. Two ferromagnetic sublattices balance one another. In (b), rows of atoms are opposed and the resulting moment along a column is zero. In (c), rows of atoms are not balanced. (d) Ferrimagnetic order. The moments of two sublattices with antiparallel alignment do not cancel one another

- ferrimagnetic, i.e., the moments of two antiferromagnetically aligned sublattices do not cancel and the resulting moment is nonzero.

In this ideal model of localised magnetism, magnetocrystalline anisotropy arises because the electrons are distributed in orbitals whose symmetry is not generally spherical, and the asymmetry of the electron cloud is strictly related to the direction of the magnetic moment. The energy arising from the electrostatic interaction between all charges in the crystal thus varies with the orientation of the magnetic moment.

However, this ideal model of localised magnetism is only very rarely observed. It is only approximately valid for ions in the rare earths, in which the  $4f$  electrons responsible for magnetic effects are located in inner shells where they are in some sense protected from the crystal field by the  $5s$  valence shell. For the other large family of magnetic atoms, the elements of the first transition series, the  $3d$  orbitals causing magnetic effects are the valence orbitals. The interaction with the crystal field can no longer be treated as a perturbation of the isolated atom and becomes even stronger than the spin-orbit interaction [3]. This reversal in the order of importance of the interactions has significant consequences, the first of which is a cancellation or quenching of the orbital moment, i.e., the quantum states arising from diagonalisation of the basic Hamiltonian (without spin-orbit coupling but with the crystal field) have zero orbital moment and the resulting magnetic moment corresponds to the spin moment alone. There is a simple classical interpretation for this phenomenon. When the effect of electrostatic interactions becomes very great, a large amount of energy must be expended to get an electron to revolve in a crystal field with spatial fluctuations, and it is therefore energetically preferable to localise it in zones of lower energy. As a consequence, the new fundamental orbitals correspond to stationary waves in the spatial probability distribution of the electrons. The localisation of the maxima of these stationary waves with respect to the spatial variations in the crystal field leads to energy differences between states.

The spin-orbit interaction is then viewed as a perturbation of this ground state and reinstates a partial orbital angular momentum. Indeed, in the presence of a spin  $\mathbf{s}$ , the reinstatement of a partial orbital angular momentum  $\delta\mathbf{l}$  in a given atomic orbital allows the system to gain in magnetic energy through the spin-orbit interaction  $H_{\text{SO}} = -\xi\delta\mathbf{l} \cdot \mathbf{s}$ . However, the cost in electrostatic energy depends on the orientation of  $\delta\mathbf{l}$ . The amplitude of the reinstated orbital angular momentum is determined by competition between these two energies and therefore depends on the orientation of the spin angular momentum with respect to the crystal lattice, as does the total energy. This is the origin of magnetocrystalline anisotropy, which in this case occurs with an anisotropy in the amplitude of the total magnetic moment. The fact that the electrostatic interaction is much stronger than the spin-orbit interaction leads to very small orbital effects, and their contribution to the total magnetic moment is often ignored. Likewise, magnetocrystalline anisotropy is generally less pronounced than in the ideal case.

### *Itinerant Magnetism*

The ferromagnetic metals, such as Fe, Co, Ni and their alloys, do not fit into the localised magnetism model. The most obvious difference is the value of the magnetic moment per atom in these metals, which cannot be explained by any filling rule for an isolated atom or ion, even by appealing to a quenching of the orbital angular momentum. In these metals, the electrons responsible for the magnetic moment are delocalised electrons, which take part in electrical conduction and must be treated using an energy band model. We then speak of itinerant magnetism.

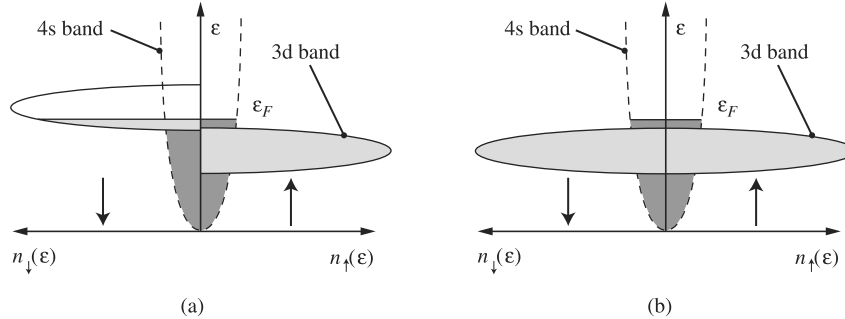
The origin of itinerant ferromagnetism can be understood quite simply using the well known Stoner model [3]. The transition metals in the group Fe, Co, and Ni are characterised by a very broad  $4s$  conduction band and a narrow  $3d$  conduction band giving a high density of states at the Fermi level in the paramagnetic state. This feature favours a polarisation of the electron spins in the  $3d$  band under the effect of the interatomic exchange interaction (see Fig. 14.3). We thus obtain a spin magnetic moment per atom proportional to the difference between the number of electrons  $n_{\uparrow}$  and  $n_{\downarrow}$  in the two spin directions, i.e.,

$$\mu_{\text{spin}} = -(n_{\uparrow} - n_{\downarrow})\mu_{\text{B}} .$$

Obviously, the magnetocrystalline anisotropy of these ferromagnetic metals is related to the generation of a partial orbital moment, as discussed above (see also [7]). Except for special cases related to nanomagnetism discussed later in the chapter, this orbital moment remains small (less than 1% of the spin moment in the case of Co, for example [8]).

In fact many experimental results concerning the  $3d$  metals can be better interpreted using a localised model rather than a purely itinerant approach, and this has given rise to heated debate. Without going into detail, a quick





**Fig. 14.3.** Schematic representation of the electron density of states  $n(\varepsilon)$  of the conduction bands (a) in a strongly ferromagnetic 3D metal (the minority spin band is full), and (b) in copper.  $\varepsilon$  is the energy and  $\varepsilon_F$  the Fermi level. The representation (a) applies fairly well to Ni and Co, whereas Fe exhibits a more complex structure. Ag and Au, other noble metals with Cu, have similar densities of states to Cu, but for the 5s (6s) and 4d (5d) bands, respectively

argument gives some idea of how the two approaches can be reconciled in a standard material. The overlap between magnetic orbitals remains small and the 3d electrons are barely delocalised compared with the 4s conduction electrons. Instead of applying a standard conduction model, one can thus consider the 3d electrons as hopping from one atom to another with an extremely rapid characteristic hopping frequency compared with other characteristic times in the system. The exchange interaction controls the hopping probability, which therefore depends on the relative orientation of the electron spin and the spin of the atoms. It thereby contributes to stabilising at a given atomic site an average value of the magnetic moment that conforms to the predictions of the itinerant model. This average value can be identified with an isolated magnetic moment at the longer characteristic times of magnetic fluctuations (spin waves, excitations destroying long-range order close to  $T_C$ ).

Finally, it should be noted that the polarisation of conduction electrons in itinerant ferromagnetic metals has extremely important consequences for electron transport, a subject discussed in Sect. 14.2.

### 14.1.3 Magnetism in Matter: Continuum Approximation

Although the approach described in Sect. 14.1.2 identifies the microscopic origins of magnetic behaviour, its implementation within a real object, even on the scale of a few nanometers, amounts to treating an  $N$ -body problem with  $N \gg 1$ . In fact, in a rather surprising way, magnetic media can be treated much more easily using the continuum approximation, right down to the atomic scale or almost, introducing minor corrections where necessary, as we shall see below.

This approximation is also based on the idea of electromagnetic forces. As in the vacuum, an electron propagating through matter is subject to the Lorentz force, but the electric field  $\mathbf{e}$  and magnetic induction  $\mathbf{b}$  fluctuate very rapidly both spatially and temporally. However, over a long distance compared with the interatomic distance, the trajectory can be worked out as if the electron were subject to macroscopic average values  $\mathbf{E}$  and  $\mathbf{B}$ . On this basis, we may neglect fluctuations on the atomic scale and calculate using average values in a continuous medium.

### Different Energies in the Continuum Approximation

We define the magnetisation  $\mathbf{M}$  as the dipole magnetic moment per unit volume (in A/m). To deal with these media, the magnetic field  $\mathbf{H}$  is easier to manipulate than the induction  $\mathbf{B}$ . In the vacuum, these two quantities are simply proportional, i.e.,  $\mathbf{B} = \mu_0 \mathbf{H}$ . In magnetised matter, the relationship is more involved:

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}), \quad \text{with} \quad \mathbf{H} = \mathbf{H}_{\text{ext}} + \mathbf{H}_{\text{D}}, \quad (14.9)$$

where  $\mathbf{H}_{\text{ext}}$  is the field produced by sources outside the magnetic system under consideration (magnetic moment and current distributions), and  $\mathbf{H}_{\text{D}}$  is the demagnetising field, i.e., the field produced in the system by its own magnetisation  $\mathbf{M}$  (see below).

To simplify expressions in what follows, we shall make two restrictive assumptions:

- Since it is generally the case in experiments, we shall assume that the field  $\mathbf{H}_{\text{ext}}$  is uniform over the relevant volume  $V$  of the sample. If this were not so, there would be a force that would tend to displace the sample in the magnetic field gradient, together with a couple on the magnetisation [3].
- We shall also assume that the sample is made from a homogeneous magnetic material, whose magnetisation has uniform magnitude  $M = M_S$  throughout the volume. We may then write  $\mathbf{M} = M_S \mathbf{m}$ , where  $\mathbf{m}$  is a unit vector with components  $m_X$  and  $m_Z$  in an orthonormal frame. It is always possible to reduce to this hypothesis by dividing the system into homogeneous regions and then introducing interface effects (interactions, inhomogeneities) and dipole interactions.

### Demagnetising Field in a Magnetic Element

The demagnetising field  $\mathbf{H}_{\text{D}}$  is the field produced in a magnetic element by its own magnetisation distribution. The analogy between electrostatics and magnetostatics mentioned in Sect. 14.1.1 can be used to calculate at any point of space the magnetic field produced by a magnetisation distribution  $\mathbf{M}$  of volume  $V$  and surface area  $S$ , by introducing a distribution of magnetic pseudo-charges with density  $\rho = -\text{div} \mathbf{M}$

and surface density  $\sigma = \mathbf{n} \cdot \mathbf{M}$ , where  $\mathbf{n}$  is the outward unit normal vector at the relevant point of the surface. It is assumed that the magnetic element in question is made from a homogeneous material whose magnetisation  $\mathbf{M}$  has uniform magnitude  $M_S$ . Hence,

$$\mathbf{M} = M_S \mathbf{m} ,$$

where  $\mathbf{m}$  is a unit vector with components  $m_X, m_Y, m_Z$ .

- Demagnetising factors of an ellipsoid. The most general case encountered in nanomagnetism is one in which  $\mathbf{m}$  has quasi-uniform orientation. Then  $\rho \approx 0$  and only surface charges contribute. Even in this case, the dipole field  $\mathbf{H}_D$  is only uniform if the element has ellipsoidal shape. It is then given by

$$\mathbf{H}_D = -N\mathbf{M} , \quad (14.10)$$

where  $N$  is a  $3 \times 3$  tensor with trace 1. Taking Cartesian coordinates along the principal axes of the ellipsoid,  $N$  is diagonal and reduces to three demagnetising factors  $N_X, N_Y$  and  $N_Z$  such that  $N_X + N_Y + N_Z = 1$ . The component  $H_{Di}$  of the demagnetising field,  $i = X, Y, Z$ , can be written

$$H_{Di} = -N_i M_S m_i .$$

For an arbitrary ellipsoid,  $\mathbf{H}_D$  depends on the orientation of the magnetisation with respect to the sample shape. This effect reflects the fundamental anisotropy of the interaction between two magnetic dipoles [see (14.4)], combined with the integration over the sample shape.

The sphere is the simplest ellipsoid. By symmetry,  $N_X = N_Y = N_Z = 1/3$  and the dipole interaction causes no anisotropy in  $\mathbf{H}_D$ . Samples intended for precise measurements or calibration are generally made spherical. Exact expressions for  $N_X, N_Y$  and  $N_Z$  in the case of an arbitrary ellipsoid can be found in [78].

- Approximately ellipsoidal shapes. Although shapes commonly encountered in nanomagnetism are not ellipsoidal, several extremely important cases are close enough to be able to make the approximation.
  - Infinitely long cylindrical needle. The needle is assumed infinitely long in the  $Z$  direction compared with its diameter. The magnetic charges on the two faces perpendicular to the  $Z$  axis are then infinitely far away, which means that  $N_Z \sim 0$  and  $N_X = N_Y \sim 1/2$ .
  - Infinitely thin film. The film can be assumed to have infinitesimal thickness in the  $Z$  direction, when compared with its lateral dimensions. This time it follows trivially that  $N_Z \sim 1$  and  $N_X \sim N_Y \sim 0$ .
- General case. For an arbitrary shape, even if the magnetisation is strictly uniform, the demagnetising field is not, and the variations in this field can subsequently induce non-uniformities in the magnetisation. For many shapes, it is nevertheless possible to define demagnetising pseudofactors evaluated from exact calculations of the dipole energy, made by assuming a uniform magnetisation aligned along the different symmetry axes of the system. The values obtained are then fitted with (14.14) below.

An important case for applications is that of a thin wafer obtained by etching a thin film. If  $L_X$  and  $L_Y$  are the lateral dimensions in the  $Oxy$  plane and  $e$  the thickness in the  $Oz$  direction, we have  $L_X \sim L_Y \gg e$ . Clearly,  $N_Z$  will be

close to 1, while  $N_X$  and  $N_Y$  will be small with  $N_X < N_Y$  if  $L_X > L_Y$ . The case of a uniformly magnetised rectangular parallelepiped has been treated analytically [79]. The demagnetising factors of a wafer with elliptical geometry can be found in [80]. It can be observed that, even for very large lateral dimensions compared with the thickness (typically  $L_X, L_Y = 100\text{--}200$  nm for  $e = 5$  nm in magnetoelectronic applications), a wafer cannot be treated as an ellipsoid for calculation of  $N_X$  and  $N_Y$ .

### *Zeeman Energy*

The interaction energy between the field  $\mathbf{H}_{\text{ext}}$  and the magnetisation distribution in the sample can be written

$$E_Z = - \iiint_V dV \mu_0 \mathbf{M} \cdot \mathbf{H}_{\text{ext}} = \iiint_V dV \varepsilon_Z, \quad (14.11)$$

where  $\varepsilon_Z = -\mu_0 M_S \mathbf{m} \cdot \mathbf{H}_{\text{ext}}$  is the Zeeman energy density per unit volume [see (14.3)].

### *Exchange Energy*

In common ferromagnetic materials, the exchange interaction tends to keep individual magnetic moments parallel. As soon as this feature of the magnetisation is destroyed, an exchange energy is created, which increases as the local rotation of  $\mathbf{m}$  is large. Since the exchange interaction is short range, affecting only nearest atomic neighbours, the exchange energy is a local quantity that can be represented by an energy density

$$\varepsilon_{\text{exch}} = A(\nabla \mathbf{m})^2, \quad (14.12)$$

where  $A$  is the exchange constant for the material. The brief notation

$$(\nabla \mathbf{m})^2 = (\nabla m_X)^2 + (\nabla m_Y)^2 + (\nabla m_Z)^2,$$

where nabla assumes its usual interpretation, is widespread in the micromagnetism literature and should not be confused with the square of the divergence of  $\mathbf{m}$ , which is written  $(\nabla \cdot \mathbf{m})^2$ .

### *Dipole Energy and Shape Anisotropy*

The dipole energy  $E_D$  arises from the interaction between the demagnetising field and the magnetisation distribution:

$$E_D = -\frac{1}{2} \iiint_V dV \mu_0 \mathbf{H}_D \cdot \mathbf{M}. \quad (14.13)$$

This has the same form as the Zeeman energy (interaction between a magnetic field and a magnetisation distribution), but with a factor of 1/2 which expresses the fact that the dipole energy is a self-energy.

For a uniformly magnetised sample with ellipsoidal shape,  $\mathbf{H}_D$  is also uniform (see p. 512) and the expression for the dipole energy density simplifies to

$$\varepsilon_D = \frac{1}{2}\mu_0 M_S^2 \mathbf{m} \cdot N \mathbf{m} = \frac{1}{2}\mu_0 M_S^2 (N_X m_X^2 + N_Y m_Y^2 + N_Z m_Z^2) . \quad (14.14)$$

This expression can often be applied to samples with non-ellipsoidal shape used in nanomagnetism. The demagnetising factors  $N_X$ ,  $N_Y$  and  $N_Z$  are then the ‘average’ values found by rigorous calculation.

$\varepsilon_D$  depends on the orientation of the magnetisation with respect to the sample shape. One speaks of shape anisotropy. Since this energy is proportional to  $M_S^2$ , it very often dominates over the other sources of magnetic anisotropy, imposing the direction of magnetisation at equilibrium in the absence of any external field. As  $\varepsilon_D$  can only be positive or zero, this easy magnetisation direction is the one with the lowest demagnetising factor. For example, the dipole interaction favours:

- alignment of the magnetisation along the length of a magnetised needle,
- maintenance of the magnetisation in the plane of a film that is very thin compared with its lateral dimensions,
- a plane of easy magnetisation in the plane of an elliptically shaped wafer, with an axis of easy magnetisation along the major axis of the ellipse.

#### *Magnetocrystalline Anisotropy*

Magnetocrystalline anisotropy refers to the energy contribution arising directly from the interaction between the magnetic moment and the symmetry of the atomic environment in the material. The origins of this magnetic energy have been described in the previous sections. It is a local energy which depends at each atomic site on the immediate atomic environment. The expressions below only consider the variable part of this energy up to a factor which can nevertheless vary itself if the crystal lattice is deformed, as we shall see later.

The commonest case is a monocrystal of a simple ferromagnetic material such as Fe, Co, or Ni. Then all atomic sites are equivalent, the energy density of magnetocrystalline anisotropy is uniform in the material, and its expansion in terms of the direction cosines of the magnetisation ( $m_X$ ,  $m_Y$  and  $m_Z$ ) must respect the symmetry of the crystal lattice and time reversal symmetry. Depending on the lattice symmetry, we use a complete basis of spherical harmonics or functions of the direction cosines  $m_i$ .

The simplest case is one of uniaxial magnetocrystalline anisotropy, where the energy density is

$$\varepsilon_{mc} = K \sin^2 \theta , \quad (14.15)$$

as a function of the angle  $\theta$  between the direction of magnetisation  $\mathbf{m}$  and the axis of anisotropy. If the anisotropy constant  $K$  is positive, the energy is

minimal when the magnetisation is parallel to the anisotropy axis (whatever direction it may be pointing). We then speak of an easy axis of magnetisation, or just an easy axis. Otherwise, the magnetisation remains preferentially in the plane perpendicular to the anisotropy axis and we speak of an easy plane and a hard axis.

For a cubic crystal,

$$\varepsilon_{\text{mc}} = K_1 s + K_2 p + K_3 s^2 + K_4 s p + \dots, \quad (14.16)$$

where  $s = m_X^2 m_Y^2 + m_Y^2 m_Z^2 + m_Z^2 m_X^2$  and  $p = m_X^2 m_Y^2 m_Z^2$ . This expansion is valid for Fe (body-centered cubic structure) and Ni (face-centered cubic structure). For a crystal with hexagonal symmetry such as Co, we have

$$\varepsilon_{\text{mc}} = K_1 \sin^2 \theta + K_2 \sin^4 \theta + K_3 \sin^6 \theta + K'_3 \sin^6 \theta \cos 6\phi + \dots, \quad (14.17)$$

where  $\theta$  is the angle between the magnetisation and the axis of hexagonal symmetry  $\mathbf{c}$ , and  $\phi$  is the angle to the azimuth in the plane perpendicular to  $\mathbf{c}$  (cylindrical coordinate system).

The coefficients  $K_i$  are called the order  $i$  anisotropy constants.

**Note.** The notation commonly used in the literature and adopted here can lead to a certain confusion. For example, the constant  $K_1$  in the expansion for cubic symmetry corresponds to a term in  $m_i^4$ , whereas the term in  $K_1$  in the expansion for hexagonal symmetry corresponds to a term in  $m_i^2$ . The term ‘order’ thus refers to the position of the term in the given expansion rather than the power of that term, and care must be taken when comparing the results of different authors.

With more involved expressions such as (14.16) and (14.17), there may exist several types of extrema in the anisotropy energy. We speak of an easy direction, where ‘direction’ may mean axis, plane, cone, etc., when the energy has a global minimum, a hard direction when the energy has a maximum, and intermediate directions when the energy has a local minimum.

Finally, the magnitude  $M_S$  also depends on the orientation of the magnetisation due to the anisotropy of the orbital moment (see Sect. 14.1.2), following the same expansion as the magnetocrystalline anisotropy with a constant term  $M_0$ .

Table 14.1 gives the coefficients in the above expansions for Fe, Co and Ni, at the temperature of liquid helium ( $T = 4.2$  K). As one might expect, the strength of anisotropy effects falls as the symmetry increases, e.g., in passing from a hexagonal to a cubic lattice, not only does the first term in the expansion correspond to a higher power of the  $m_i$ , but the value of the coefficient is smaller. The value of the coefficients  $K_i$  generally falls off rather quickly as the order increases, and it is justifiable to keep only the first one or two terms. Care must also be taken because the terms of higher order, although

**Table 14.1.** The main parameters of magnetic anisotropy for ferromagnetic transition metals at  $T = 4.2$  K [9]

	Fe	Co	Ni
Crystal lattice	cc, $a = 2.87 \text{ \AA}$	hcp, $b = 2.51 \text{ \AA}$ , $c = 4.07 \text{ \AA}$	fcc, $a = 3.52 \text{ \AA}$
Anisotropy energy			
$K_1$ [eV/atom]	$4.02 \times 10^{-6}$	$5.33^{-5}$	$-8.63 \times 10^{-6}$
$K_2$ [eV/atom]	$1.44 \times 10^{-8}$	$7.31 \times 10^{-6}$	$3.95 \times 10^{-6}$
$K_3$ [eV/atom]	$6.6 \times 10^{-9}$		$2.38 \times 10^{-7}$
$K'_3$ [eV/atom]		$8.4 \times 10^{-7}$	
Moment anisotropy			
$M_0$ [ $\mu_B$ /atom]	2.215	1.729	0.615
$M_1$ [ $\mu_B$ /atom]	$-5.4 \times 10^{-4}$	$-8.0 \times 10^{-3}$	$6.0 \times 10^{-4}$

small, may make a significant contribution in certain experiments. Finally, the anisotropy of the moment is extremely small and remains negligible in almost all experiments.

#### *Magnetoelastic Anisotropy*

The above discussion referred to perfect lattices. Magnetic nanostructures are rarely independent of their environment and even in a self-supported structure (a bulk cluster), size effects generally lead to distortion or relaxation of the crystal structure, as for a surface or interface. The anisotropy depends on the immediate atomic environment and so any deformation of the lattice can change the form of the expansion and the coefficients of magnetocrystalline anisotropy.

We speak of magnetoelastic anisotropy when this variation corresponds to the elastic deformation of the material in response to a stress. The change depends sensitively on the initial symmetry of the lattice. For example, a stress applied along the axis  $c$  of a lattice with hexagonal symmetry does not change the form of the first terms in the expansion, but simply alters the values of the coefficients. In contrast, a stress exerted along one of the axes of symmetry of a cubic lattice induces a tetragonal distortion which introduces higher order terms in the expansion of  $\varepsilon_{mc}$ , viz.,

$$\varepsilon_{mc} = K_1 \sin^2 \theta + K_2 \sin^4 \theta + K'_2 \sin^4 \theta \cos 4\phi + \dots, \quad (14.18)$$

where  $\theta$  and  $\phi$  are the same angles as for the hexagonal symmetry [see (14.17)] but relative to the axis of tetragonal symmetry.

#### *Anisotropy Due to Bond Direction*

Up to now we have been considering single-element materials. In an alloy there may exist another type of anisotropy due to the direction of the bonds

between the various elements in the alloy. A classic case is provided by ordered alloys with  $L1_0$  structure such as  $\text{Fe}_{50}\text{Pt}_{50}$ , which is the alloy with the highest anisotropy coefficient yet observed, apart from the rare earth alloys. In this ordered structure, the atoms are situated on a face-centered cubic (fcc) structure, with pure planes of each constituent Fe or Pt alternating along a [001] direction. The anisotropy thus accumulates the effects of an anisotropy in the bond direction and a high level of tetragonal distortion of the lattice due to the great difference in volume between Fe and Pt atoms. To this must be added the favourable influence of the very strong spin-orbit coupling of Pt (a heavy atom), rendered magnetic by contact with Co (see Sect. 14.1.4).

The effect is more subtle in the layers of alloys of type  $\text{Ni}_{1-x}\text{Fe}_x$ , in which a deposition or anneal in a magnetic field can induce by diffusion a spatial organisation of the Ni-Fe, Fe-Fe and Ni-Ni bonds in such a way that more bonds of a given type are aligned close to the field direction. This effect then induces a weak magnetic anisotropy which can be extremely useful in magnetoelectronic applications.

### *Magnetostriction*

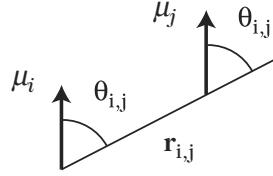
The relationship between crystal structure and magnetisation is a much more complicated affair in which the phenomena of magnetocrystalline anisotropy discussed above represents only one aspect. In fact, it is the global energy which must be minimised when the magnetisation rotates in a magnetic material: not only does the amplitude of the magnetic moment vary (see Sect. 14.1.2), but the crystal lattice may be deformed. This last phenomenon is called magnetostriction, characterised in the experimental context by a variation in the dimensions of a macroscopic magnetic sample depending on the direction of magnetisation. This phenomenon is extremely important for magnetoelectronic applications, because the behaviour of the magnetisation of a ferromagnetic layer in a device can be significantly influenced by the interaction between magnetostrictive effects and mechanical stresses exerted in the neighbourhood of the layer. Hence, most applications requiring rotationally free layers use permalloy layers, i.e.,  $\text{Ni}_{80}\text{Fe}_{20}$  alloy, whose magnetostriction can be cancelled by adjusting the concentration. However, other applications use materials with high magnetostriction to detect a mechanical stress, e.g., sensors, keyboards, etc.

It is not possible to go into more detail here. For the physicist who wishes to understand these effects qualitatively, the pair model proposed by L. Néel [10] provides an excellent unifying tool with regard to the various types of magnetic energy connected with the crystal lattice and its deformations (see below).

### **Néel Pair Model**

We shall just outline the basic idea behind this method, introduced by L. Néel in 1953 to calculate the surface anisotropy [10]. We first assume that the exchange





**Fig. 14.4.** Notation for pair model

interaction is strong enough to keep all magnetic moments parallel in the crystal. The total magnetocrystalline energy can then be written in the form of a sum of the pair energy  $e_{ij}$  over all crystal sites, viz.,

$$E_{\text{crystal}} = \frac{1}{2} \sum_{i,j} e_{ij} .$$

With the parameters defined in Fig. 14.4, the energy  $e_{ij}$  of a pair of moments  $\mu_i$  and  $\mu_j$  can be expanded without loss of symmetry using a basis of Legendre polynomials:

$$e_{ij} = g_2(r_{ij})P_2(\theta_{ij}) + g_4(r_{ij})P_4(\theta_{ij}) + \dots .$$

The functions  $g_n(r_{ij})$  are assumed to decrease very quickly with  $r_{ij}$ . At the lowest level of approximation, the sum is therefore taken over nearest neighbours, keeping only the first term in the expansion of  $e_{ij}$ . Furthermore,

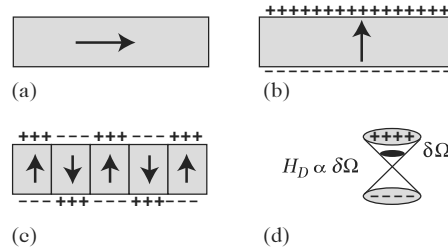
$$P_2(\theta_{ij}) = \cos^2 \theta_{ij} - \frac{1}{3} , \quad g_2(r_{ij}) = a + b\delta r_{ij} , \quad \text{with } r_{ij} = r_0 + \delta r_{ij} .$$

Note that  $P_2$  has the symmetry of the magnetic dipole interaction between two parallel dipoles, although this interaction is generally treated separately. The expression for  $g_2$  represents the elastic energy due to deformations of the crystal lattice about its equilibrium position  $r_0$ .

In principle, these expressions can be used to unify the description of magnetocrystalline anisotropy with its magnetostriction and magnetoelastic energy aspects. For example, the determination of the coefficients  $a$  and  $b$  from measurements of magnetostriction and elastic deformation should allow evaluation of the magnetic anisotropy coefficients. The calculation can also be extended to alloys to obtain the anisotropy due to bond directions. In practice, the method does not lead to reliable quantitative predictions. For one thing, it is very sensitive to the number of neighbours taken into account and the number of terms used in the expansion. However, it remains extremely useful for assessing the symmetry of the first terms in the expansion of the magnetic anisotropy energy in special cases of symmetry breaking often encountered in nanomagnetism [81].

## Fundamental Structures and Lengths in Nanomagnetism

The magnetisation configuration in a magnetic element is the direct result of competition between the different energy sources described above. It is usually

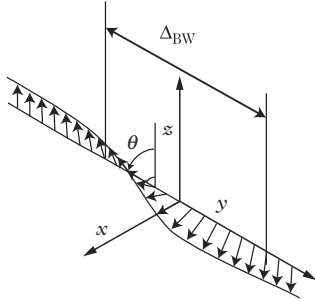


**Fig. 14.5.** Magnetic configurations in a thin film with uniaxial magnetocrystalline anisotropy and easy axis perpendicular to the film. *Arrows* indicate the orientation of the magnetisation, whilst the *plus and minus signs* indicate the magnetic surface charges. **(a)** Uniform magnetisation in the plane of the film. **(b)** Uniform magnetisation perpendicular to the film. **(c)** Magnetic domains. **(d)** At the domain centre, the dipole field is proportional to the solid angle subtended there by the magnetic pseudo-charges

the exchange interaction that dominates, despite its very short range. The competition between the exchange energy and other energy sources will thus determine, for example, the minimal distances over which the magnetisation can rotate, and these distances play the role of fundamental lengths defining the scale of nanomagnetic effects.

#### *Magnetic Domains and Walls*

The first consequence of this competition in sufficiently large structures is the spontaneous formation of magnetic domains over which the magnetisation is uniform. These regions are separated by domain walls. The origin of this effect is easily understood in the classic example investigated by C. Kittel in 1946 [11], viz., the case of a thin magnetic film with uniaxial magnetocrystalline anisotropy in which the easy axis is perpendicular to the film. The exchange energy seeks to maintain a uniform magnetisation throughout the film. The magnetocrystalline anisotropy is minimal if the magnetisation is perpendicular to the film, but this orientation gives a maximal dipole energy (see Figs. 14.5a and b). The film thus divides spontaneously into magnetic domains with limited lateral extension (see Fig. 14.5c). In a given domain, the magnetisation is indeed uniform and oriented perpendicularly to the film, while the dipole interaction decreases as the lateral extent of the domain decreases (see Fig. 14.5d). However, in the walls separating the domains, the magnetisation must rotate and this rotation has a cost in terms of exchange energy and magnetocrystalline anisotropy energy, whose minimisation confers a finite width on the wall. The competition between the wall energy (proportional to the total surface area of the wall) and the dipole interaction then determines the lateral extent of the domains.



**Fig. 14.6.** Bloch wall between two domains with antiparallel magnetisation in the  $z$  direction. In the absence of perturbation, the wall is planar, assumed here to be parallel to the  $xz$  plane. Crossing the wall in the perpendicular direction  $y$ , the magnetisation rotates steadily, whilst remaining all the time in the  $xz$  plane

#### *Competition Between Exchange Energy and Magnetocrystalline Anisotropy Energy*

The perfect Bloch wall shown in Fig. 14.6 is the magnetisation structure with minimal energy making the transition from one domain of uniform magnetisation to another with opposite magnetisation, in a magnetic material with uniaxial magnetocrystalline anisotropy and infinite extent. In this case, there is no dipole anisotropy related to the shape of the material, and the rotation of the magnetisation in the wall leads to  $\text{div} \mathbf{m} = 0$ . The width  $\Delta_{\text{BW}}$  of the Bloch wall is thus the fundamental length measuring competition between exchange energy and magnetocrystalline anisotropy energy:

$$\Delta_{\text{BW}} = 2\sqrt{\frac{A}{K}}. \quad (14.19)$$

Likewise, the expression for the energy density  $\sigma$  of the wall per unit area only involves the exchange and magnetocrystalline anisotropy coefficients:

$$\sigma = 4\sqrt{AK}. \quad (14.20)$$

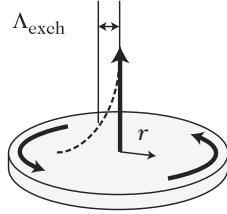
For typical materials,  $\Delta_{\text{BW}}$  ranges over 7–100 nm.

#### *Competition Between Exchange Energy and Dipole Interaction*

When the total magnetic energy in a nanostructure is minimised, the resulting magnetisation configuration may not be uniform over characteristic length scales determined by the competition between energies. The competition between the exchange and dipole energies determines the exchange length  $\Lambda_{\text{exch}}$  given by

$$\Lambda_{\text{exch}} = \sqrt{\frac{2A}{\mu_0 M_S^2}}. \quad (14.21)$$

The simplest way of visualising this length is to consider the magnetisation vortex (see Fig. 14.7), as observed, for example, in a circular disk of a thin



**Fig. 14.7.** Magnetisation vortex in a magnetic disk. *Bold arrows* indicate the local magnetisation direction in the material and  $r$  is the distance to the disk centre

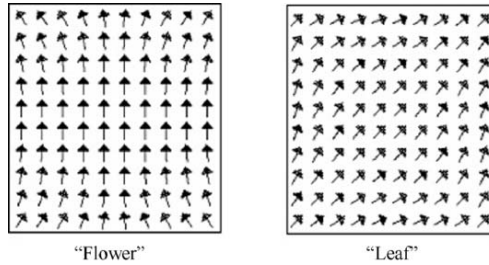
film of magnetic material with negligible magnetocrystalline anisotropy (a soft material). To minimise the dipole energy in such a disk, the magnetisation must remain in the plane so that there are no magnetic surface charges on the upper and lower faces, and it must be parallel to the edge so that there are no magnetic charges on the circumference. The circular symmetry of the magnetisation configuration obtained in this way also suppresses bulk charges ( $\text{div}\mathbf{m} = 0$ ). However, it does lead to a singularity at the centre of the disk which the magnetisation deals with by coming out of the plane. The profile of the perpendicular component of  $\mathbf{m}$  is then given by the expression

$$m_{\perp} = \exp \left[ - \left( \frac{r}{\Lambda_{\text{exch}}} \right)^2 \right], \quad (14.22)$$

which involves the exchange length. Whether a vortex is actually obtained clearly depends on the disk parameters, e.g., magnetisation  $M_S$ , diameter, thickness. Intuitively, one may predict that the vortex structure becomes unstable in a very thin film, where the energy arising from surface charges becomes relatively large, or when the diameter is reduced [12, 13].

For typical materials,  $\Lambda_{\text{exch}}$  has small values, of the order of 3–5 nm, well below  $\Delta_{\text{BW}}$ . This simply reflects the fact that the dipole interaction generally dominates the magnetocrystalline anisotropy.

The exchange length actually measures the ability of the magnetisation to rotate over very short distances to minimise its total energy by reducing its dipole energy at the expense of the exchange energy. Hence, in magnetic objects with non-ellipsoidal shapes, where the demagnetising field is not uniform, the magnetisation can only be strictly uniform when the dimensions are of the order of  $\Lambda_{\text{exch}}$ . A clear illustration is given in Fig. 14.8, which gives a schematic view of the magnetisation configurations in a square cell etched in a soft magnetic film. Near the corners of the square and along the edges perpendicular to the magnetisation, the demagnetising field would be very strong if the magnetisation remained uniform. It thus rotates in order to minimise the magnetic surface charges, a process occurring over a distance of the order of  $\Lambda_{\text{exch}}$ . Note that the two configurations in Fig. 14.8 do not have the same total energy. This is called a configuration anisotropy energy [14]. These non-uniformities on the scale of  $\Lambda_{\text{exch}}$  also have considerable consequences for the way the magnetisation of a nanostructure rotates under the effect of an applied field, a critical process in applications [15, 16].



**Fig. 14.8.** Quasi-uniform magnetisation configurations minimising the magnetic energy in a square cell etched in a thin film of soft magnetic material. *Left:* Flower. *Right:* Leaf [14]

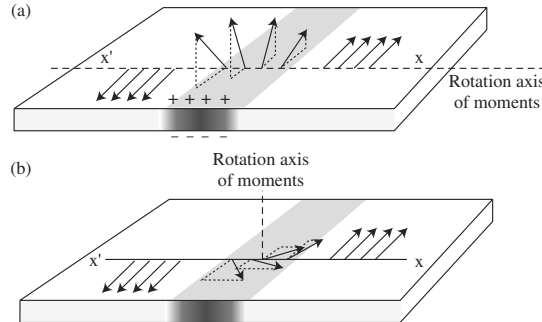
#### *Magnetic Domain Walls in Finite-Size Samples*

When the lateral dimensions of a magnetic structure become comparable with the fundamental lengths discussed above, or over a depth of the order of these lengths close to the surface of a more extended structure, highly non-uniform magnetic configurations are encountered. The structure shown in Fig. 14.8 is an example. The reader is referred to the exhaustive treatment published by A. Hubert and W. Rave [17]. The same authors have also discussed in detail a classic example, the cube [18], while H. van den Berg has developed a general method [19] for predicting the magnetic configurations in elements etched in soft ferromagnetic films.

On the level of the present discussion, it is possible to obtain a simple understanding of the basic principles of these complex magnetic configurations from a brief discussion of a case that is extremely important for applications, namely, magnetic domain walls in thin films.

The simplest configuration is the one in an ultrathin film, defined as a film whose thickness is at most of the order of the exchange length. Consequently, exchange succeeds in keeping the magnetisation parallel over the thickness of the film. Let us assume that the magnetocrystalline anisotropy creates an easy magnetisation axis in the plane of the film. The film can then divide up into magnetic domains that are uniformly magnetised along the easy axis, as shown schematically in Fig. 14.9. Two minimum energy configurations are then possible for the walls:

- The magnetisation can rotate as in a Bloch wall, hence coming out of the plane (Fig. 14.9a). This configuration creates magnetic surface pseudo-charges which cost a certain amount of dipole energy. Note that, due to these charges, this is not exactly a perfect Bloch wall. However, when the width of the wall is much greater than the film thickness, it can be shown that the width and energy of the wall are given approximately by (14.19) and (14.20), replacing the magnetocrystalline anisotropy constant  $K$  by



**Fig. 14.9.** Domain walls in an ultrathin film. (a) Bloch-type wall. (b) Néel wall

the effective constant  $K_{\text{eff}} = K - 2\pi M_S^2$ , which includes the dipolar shape anisotropy of a thin film [see (14.14) and p. 512].

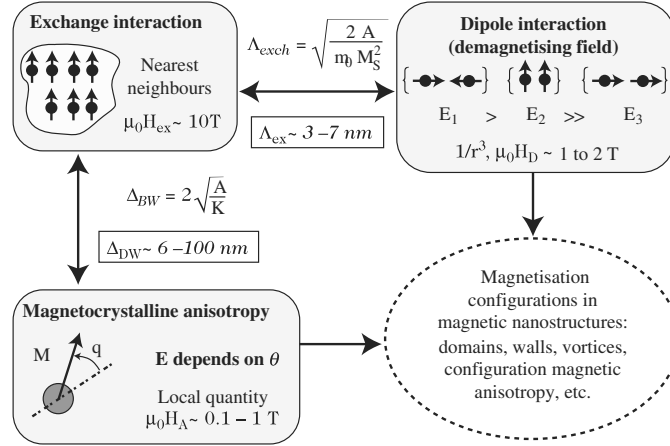
- The magnetisation can also rotate within the plane. This last structure is called a Néel wall. Since the magnetisation does not leave the plane, there are no surface pseudo-charges, in contrast to the first case. However, in the wall,  $\text{div} \mathbf{m} \neq 0$  and there are therefore bulk pseudo-charges. The Néel wall has a complex structure [20]. It has a core whose width is of order  $\Lambda_{\text{exch}}$ , in which the main part of the rotation of the magnetisation occurs, surrounded by extended zones in which the rotation is slower. The extent of the latter is proportional to the film thickness and the ratio  $\mu_0 M_S^2 / K$ .

Without actually carrying out the minimisation calculation, it can be seen that the cost in dipole energy is related to surface effects in the first case and proportional to the volume of the wall in the second case. The Néel wall is thus favoured below a certain critical film thickness which depends on the magnetic constants of the material.

When the film thickness increases well beyond  $\Lambda_{\text{exch}}$ , the exchange interaction is no longer able to maintain a parallel magnetisation throughout the whole thickness of the film. In order to minimise all the energies, the wall will adopt a Néel-type configuration close to the surfaces and a Bloch-type configuration at the film centre [17, 21].

When the lateral dimensions of the film are limited, even more complex situations are observed. For example, in a narrow strip of magnetic material, the shape anisotropy tends to hold the magnetisation parallel to the edges of the strip, whilst the wall will align itself perpendicularly to the strip to minimise its length. This kind of wall, perpendicular to the magnetisation is called a head-to-tail wall. It assumes complex structures that are very sensitive to the dimensions of the strip [22].

Figure 14.10 sums up all the different points discussed in this section, as a reference for understanding the magnetisation configuration in magnetic micro- and nanostructures.



**Fig. 14.10.** Competition between different forms of magnetic energy. Fundamental lengths of nanomagnetism

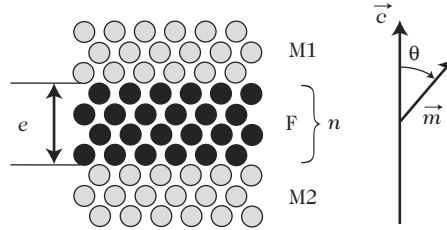
#### 14.1.4 Novel Magnetic Effects on the Nanoscale

When at least one dimension of an object becomes of the order of a few interatomic distances, major differences are expected in the equilibrium states compared with the properties of the bulk material:

- The stable crystal structure can be different, in particular, under the effects of interaction with the environment.
- By straightforward geometric effects, the properties of surface and interface atoms take on an overwhelming significance in determining the properties of the object as a whole.
- The confinement of particles such as electrons or magnons within distances comparable with their quantum wavelength induces novel behaviour, known as quantum size effects.

The first effect is not strictly a magnetic effect, even though the interaction with a substrate allows one to stabilise and study new phases of magnetic materials in ultrathin films, e.g., the bcc structure of Co or the hcp structure of Fe.

The two other effects have made it possible to observe many novel types of magnetic behaviour which we shall illustrate by several important examples studied on ultrathin films. Indeed, these films represent powerful tools for measuring interface and confinement effects, and they are extensively used in the magnetoelectronic applications discussed below. Note that effects of the same kind are found in clusters, discussed in Chap. 7.



**Fig. 14.11.** Ferromagnetic film F comprising  $n$  atomic planes sandwiched between two non-magnetic films  $M_1$  and  $M_2$

### Interface Magnetic Anisotropy

The magnetic anisotropy of ultrathin films is not only a classic example for illustrating interface effects, but also a practical model for expressing the influence of defects occurring in real films.

#### *Interface Magnetocrystalline Anisotropy*

An interface between two materials (or a surface) is a place where translation symmetry is rather suddenly broken, and where the overlap between atomic orbitals of different atoms can significantly modify the local structure of electronic bands. Now magnetic anisotropy is a local quantity which depends on the crystal symmetry in the immediate vicinity of an atom. In 1953, L. Néel predicted [10] that the atoms located on a surface or interface would possess very different magnetic anisotropy coefficients to atoms in the bulk material.

The phenomenon has been widely studied in Co films and multilayers, owing to the importance of such effects for magnetic recording applications. This case also corresponds to a model calculation of the interface effect which we shall now examine.

Consider an ultrathin film of Co with hcp structure and hexagonal symmetry axis  $c$  perpendicular to the film. Suppose the film has a thickness of  $n$  atomic planes, sandwiched between two non-magnetic materials  $M_1$  and  $M_2$  with the same structure and semi-infinite extent (see Fig. 14.11). All the Co atoms in this structure have an atomic environment with symmetry axis perpendicular to the film. Hence, to first order, the anisotropy energy per atom can be written as for uniaxial anisotropy, viz.,  $\varepsilon = k \sin^2 \theta$ . To all intents and purposes, the anisotropy coefficient  $k$  per atom depends only on the nearest neighbours of the atom. There are thus only two types of Co atom in this context: those inside the film and those in the interfacial atomic planes. These are denoted by  $k_{\text{bulk}}$ ,  $k_{I_1}$  and  $k_{I_2}$ . Moreover, if the film is thin enough, e.g., compared with  $\Delta_{\text{BW}}$  or  $\Lambda_{\text{exch}}$ , all the magnetic moments are held parallel by the exchange interaction and the total magnetic anisotropy can be written as a simple average over the atoms along a direction perpendicular to the film:

$$\varepsilon_{\text{mc}} = \frac{\rho}{n} \left[ (n-2) k_{\text{bulk}} + k_{I_1} + k_{I_2} \right] \sin^2 \theta, \quad (14.23)$$



where  $\rho$  is the atomic density, assumed here to be uniform over the stack. We may then return to the continuum approximation by introducing in the bulk of the film the bulk and interface anisotropy coefficients  $K_{\text{bulk}}$ ,  $K_{S_1}$  and  $K_{S_2}$  corresponding to

$$K_{\text{bulk}} = k_{\text{bulk}}\rho, \quad K_{S_i} = \rho d(k_{I_i} - k_{\text{bulk}}), \quad (14.24)$$

where  $d$  is the distance between atomic planes parallel to the interface. With film thickness  $e = nd$ , the magnetocrystalline energy density per unit volume becomes

$$\varepsilon_{\text{mc}} = \left( K_{\text{bulk}} + \frac{K_{S_1} + K_{S_2}}{e} \right) \sin^2 \theta. \quad (14.25)$$

To deal with the global anisotropy of the film, we introduce the dipolar shape anisotropy. In the limit as the film becomes very thin, this simplifies to [see (14.14) and p. 512]

$$\varepsilon_{\text{D}} = -\frac{\mu_0}{2} M_{\text{S}}^2 \sin^2 \theta. \quad (14.26)$$

Finally, we obtain a uniaxial magnetocrystalline anisotropy with effective coefficient

$$K_{\text{eff}} = K_{\text{bulk}} - \frac{\mu_0}{2} M_{\text{S}}^2 + \frac{K_{S_1} + K_{S_2}}{e}. \quad (14.27)$$

Real films rarely have the ideal structure in Fig. 14.11, due to interface roughness and interdiffusion. However, (14.27) often provides an excellent description of their properties (see Appendix A at the end of the chapter). To check the relation,  $eK_{\text{eff}}$  is plotted as a function of  $e$ , as in Fig. 14.12. The slope of the straight line then gives the factor

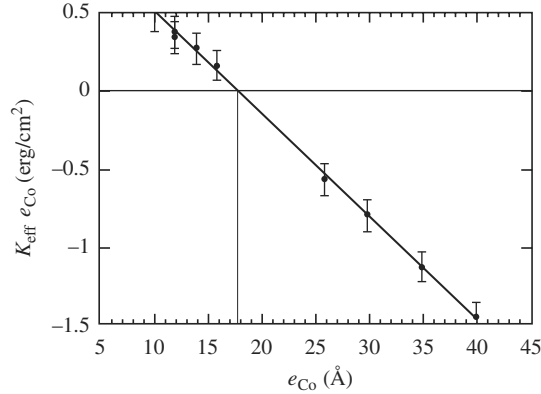
$$K_{\text{bulk}} - \frac{\mu_0}{2} M_{\text{S}}^2,$$

and the point of intersection with the vertical axis gives

$$K_{S_1} + K_{S_2}.$$

The case illustrated in Fig. 14.12 concerns an Au/Co/Au(111) sandwich [23].

Equation (14.27) leads to an extremely interesting result. When there is no interface anisotropy, the dipole anisotropy is generally stronger than the magnetocrystalline anisotropy and maintains the magnetisation in the plane of the film ( $K_{\text{eff}} < 0$ ). For example, with the coefficients for bulk Co at room temperature, we have  $K_{\text{bulk}} = 0.45 \times 10^6 \text{ J/m}^3$  and  $\mu_0 M_{\text{S}}^2/2 = 1.25 \times 10^6 \text{ J/m}^3$ . However, if the interface anisotropy  $K_{S_1} + K_{S_2}$  is positive, then below a critical thickness  $e^*$ , the magnetocrystalline anisotropy causes a magnetisation along a perpendicular easy axis. This result was first observed in 1968 for a monolayer



**Fig. 14.12.** Characteristic plot of  $eK_{\text{eff}}$  as a function of  $e$  for an interface anisotropy in an Au/Co/Au(111) sandwich.  $e_{\text{Co}}$  = thickness of the Co film,  $K_{\text{eff}}$  = effective measured anisotropy coefficient [23]

of NiFe on Cu [24], but much larger values of  $e^*$  were then measured [25]. This effect is clearly visible in Fig. 14.12, where  $K_{\text{eff}}$  for an Au/Co/Au(111) sandwich becomes positive below a Co thickness of the order of 1.8 nm [23]. Layers of Co on Pt provided a particularly interesting example, in which hybridisation at the interface associates the strong spin-orbit coupling of the Pt with the spin polarisation of the Co, inducing a spin polarisation in the interfacial Pt atoms by a proximity effect. The Co/Pt interface anisotropy thus generates a critical thickness of 3 nm in Pt/Co/Pt(111) sandwiches [26].

More generally, through hybridisation of electronic bands between interface atoms,  $K_{\text{S}}$  acquires a sensitivity to electron states in the non-magnetic film M. Any evolution of these states, e.g., at very low thicknesses of M (localised states, quantum size effects), can lead to large fluctuations in  $K_{\text{S}}$  [27].

Finally, the form of the expansion of the anisotropy energy depends sensitively on the crystal orientation of the interface (e.g., see [28] for the Co/Pt interface), and must be determined from symmetry considerations. The Néel pair model can also be used (see p. 518). More complex effects can shift the easy magnetisation axis, involving higher orders in the expansion, and sometimes affecting axes within the plane of the film. These effects are referred to as reorientation transitions.

**Note.** In all cases, a dependence of the coefficients on  $1/e$  is the signature of interface effects in ultrathin films. Such a dependence is observed for many other properties, such as the magnetic moment, optical and magneto-optical effects, and in fact any phenomenon representing a mean value over the film thickness where the atomic coefficients vary with the immediate crystallographic environment.

*Interface Magnetoelastic Anisotropy*

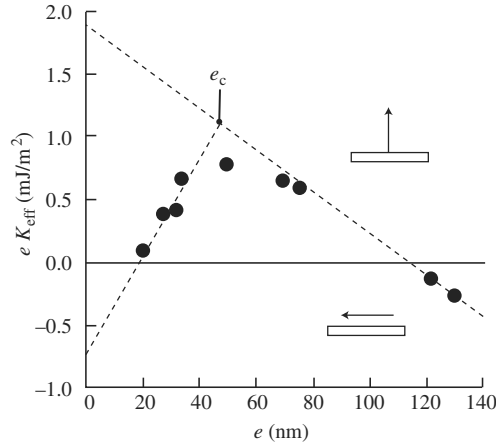
An interface between two metals is a place where two crystal lattices, which have at best the same structure but slightly different lattice constants, must adapt to one another. This matching problem induces stresses in the magnetic film and hence a magnetoelastic contribution to the magnetocrystalline anisotropy. These effects were predicted in 1988 [29] in the context of a very simple model calculation that can be usefully outlined here.

Suppose that a ferromagnetic film of metal F is deposited on a substrate S with the same crystal structure but with different lattice constant  $a_S$  from the lattice constant  $a_F$  of F. To a first approximation, it may be assumed that the crystal structure of the very thick substrate does not change during growth of the film. On the other hand, the lattice constant of F in the plane of the film, denoted by  $a$ , will depend on the thickness  $e_F$  of the film. It can be characterised by the strain  $\varepsilon = (a - a_F)/a_F$ .

In the first stages of growth, the film F adapts its lattice constant in the plane to the value for S. We then have pseudomorphic growth, and the strain remains constant at  $\eta = (a_F - a_S)/a_F$ . It causes a relaxation in the lattice constant perpendicular to the plane, according to the laws of elasticity, and hence a tetragonal distortion (as under the effect of a uniaxial stress) which costs a certain amount of bulk energy, since the film as a whole is under stress.

Beyond a critical thickness  $e_c$ , it becomes preferable to gradually relax the stress by forming dislocations at the F/S interface, in order to exchange bulk elastic energy for interface energy. If the tetragonal distortion remains elastic, it can be shown that the strain then varies approximately as  $\varepsilon \approx \eta(e_c/e_F)$ , i.e., going as the reciprocal of the film thickness. Now, still within the elastic limit, the magnetoelastic contribution to the anisotropy energy is proportional to  $\varepsilon$ . Using the formalism of (14.25), we may thus say that, up to  $e_F = e_c$ , the magnetoelastic energy adds a constant term, analogous to a bulk contribution  $K_{\text{bulk}}$ , whereas beyond  $e_c$ , it adds a contribution with the same functional dependence as an interface anisotropy, which can be included in  $K_S$ . This last term has thus been called the interface magnetoelastic anisotropy, even though it refers to the whole bulk of the film.

In real films, competition between magnetocrystalline interface anisotropy and magnetoelastic effects can lead to a complex relationship, the archetypal example of which is observed in the Cu/Ni/Cu(001) system. Figure 14.13 shows the experimental dependence of the product  $eK_{\text{eff}}$  on the thickness  $e$  of the Ni film [30]. Due to the small lattice mismatch ( $-2.5\%$ ) between the Ni and Cu layers, both of which have the fcc structure, the Ni film remains pseudomorphic up to a critical thickness  $e_c$  of the order of 46 nm, at which value  $eK_{\text{eff}}$  has a clear positive maximum. Below  $e_c$ , the positive slope is dominated by a large constant magnetoelastic contribution which induces a perpendicular easy magnetisation axis, despite the negative values of the magnetocrystalline interface anisotropy (point of intersection with the vertical axis) and the bulk term  $K_{\text{bulk}} - (\mu_0/2)M_S^2$  (negative slope for  $e > e_c$ ).



**Fig. 14.13.** Characteristic functional dependence of  $eK_{\text{eff}}$  on  $e$  for an interface magnetoelastic anisotropy in a Cu/Ni/Cu(001) sandwich, where  $e$  is the thickness of the Ni film,  $K_{\text{eff}}$  is the effective measured anisotropy coefficient,  $e_c$  is the critical magnetoelastic thickness [30]

Beyond  $e_c$ , the appearance of interface dislocations transforms the constant magnetoelastic contribution into a magnetoelastic surface term (large positive intersection with the vertical axis), which allows the easy magnetisation axis to remain perpendicular up to a record Ni thickness of 114 nm.

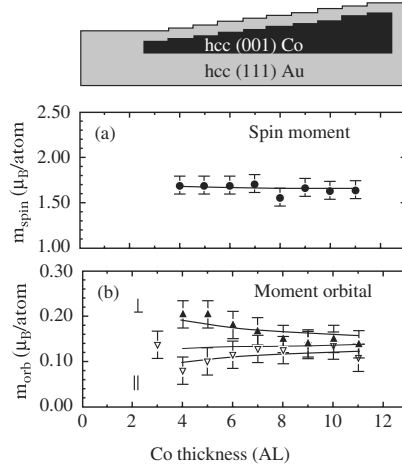
#### *Anisotropy of the Orbital Moment: From Ultrathin Films to Isolated Atoms*

In itinerant ferromagnetism, the magnetic moment  $\mu_{\text{orb}}$  of orbital origin is in principle quenched (i.e., held at zero) by the crystal field in the ground state, and it is the spin-orbit interaction, acting as a perturbation, which reattributes a small value to it. In a calculation for an isolated monatomic film, P. Bruno [7] has shown that, to a first approximation,

$$\Delta\varepsilon_{\text{mc}} = \varepsilon_{\text{mc}\perp} - \varepsilon_{\text{mc}\parallel} = - \left[ \frac{G}{H} \right] \frac{\xi}{4\mu_{\text{B}}} (\mu_{\text{orb}\perp} - \mu_{\text{orb}\parallel}), \quad (14.28)$$

where the symbols  $\perp$  and  $\parallel$  denote the values of the quantities for an orientation of the total magnetic moment that is perpendicular or parallel to the film, respectively. The symbol  $[G/H]$  denotes a numerical factor of order 1 which depends on the electronic band structure of the material. The other variables here have already been defined.

This relation expresses the fact that the anisotropy of the orbital moment is proportional to the anisotropy of the magnetocrystalline energy. Qualitatively, at a surface, the reduced symmetry of the atomic environment decreases the crystal field, which also becomes highly anisotropic. The unquenched orbital moment is thus larger, and it ends up inducing a very strong magnetocrystalline anisotropy.



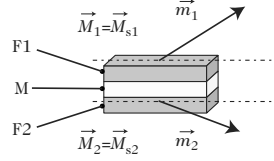
**Fig. 14.14.** On an ultrathin Au/Co/Au(111) film deposited in steps (*top view*: each step has the same Co thickness), the circular dichroism of X rays can be used to measure the perpendicular and parallel components of (a) the spin magnetic moment and (b) the orbital magnetic moment as a function of the Co thickness [32]

The relation can be extended to an ultrathin film by averaging  $\varepsilon_{\text{mc}}$  and  $\mu_{\text{orb}}$  over the film thickness. Furthermore, assuming that the factor  $[G/H]$  remains approximately constant, it is found that  $\mu_{\text{orb}}$  depends on  $1/e$  in the way described by (14.25).

The first direct check was obtained on Au/Co/Au(111) films identical to those in Fig. 14.12, using the circular dichroism of X rays to make independent measurements of the spin and orbital magnetic moments as a function of the Co thickness  $e$  [31]. Although the spin moment remains constant to within experimental error, the anisotropy of the orbital moment clearly exhibits the same dependence on  $1/e$  as the magnetocrystalline anisotropy constant  $K$  [32]. In fact, as can be seen from Fig. 14.14, each component of  $\mu_{\text{orb}}$  goes as a power of  $1/e$ , the parallel component decreasing while the perpendicular component increases.

The development of more and more sensitive experimental techniques that can be used in ultrahigh vacuum during film growth has recently made it possible to extend these investigations to still more anisotropic nanostructures, such as chains of Co atoms on a vicinal Pt(111) surface [33], or islands containing a few Co atoms on a Pt(111) surface [34]. The perpendicular orbital moment reaches record values, from  $0.3\mu_B$  for a monatomic plane and  $0.68\mu_B$  for an atomic chain, up to  $1.1\mu_B$  for an isolated adatom. Equation (14.28) continues to hold to a certain level of approximation.

These results can clearly be extended to clusters. Generally, they can show the way to making the kind of highly anisotropic magnetic nanostructures that will be required for future magnetic recording systems.



**Fig. 14.15.** Three-layer model for studying the oscillating interaction between magnetic layers through a non-magnetic metallic layer

### Electronic Quantum Confinement Effects

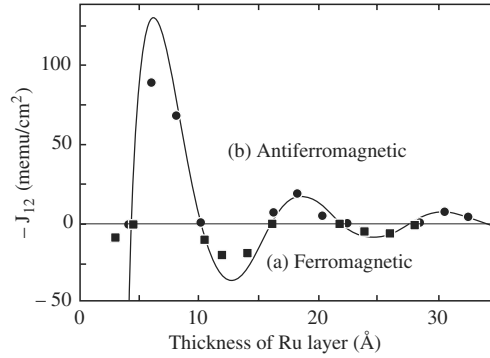
It comes as no surprise to observe interference effects in optics and they are commonly put to use in applications. For example, a Fabry–Perot interference filter confines photons between two parallel semi-reflecting mirrors. The optical transmittance at normal incidence has a peak at wavelengths that are multiples of the length  $\lambda_0$  equal to twice the distance between the mirrors. This is the condition for the electromagnetic wave associated with the photons to undergo a phase shift that is a multiple of  $2\pi$  (constructive interference) during a return trip in the filter. To observe these effects, the surface roughness of the mirrors must be much less than  $\lambda_0$ , an easy condition to fulfill in the case of visible light ( $\lambda_0 \sim 0.4\text{--}0.8\mu\text{m}$ ), and  $\lambda_0$  must be much less than the phase coherence length of the photons, something that can be achieved using lasers, for example.

A conducting multilayer is the electronic analogue of the Fabry–Perot device, and the electrons have an associated wavelength  $\lambda_e$ . One thus expects to observe electronic quantum interference effects. However,  $\lambda_e$  is of nanometric order, so an electron Fabry–Perot device must be almost perfect on the atomic scale. Electronic quantum effects were nevertheless observed early on in semiconducting multilayers, whose structural qualities can be almost perfect, and they are now used in optoelectronic devices.

In 1990, S. Parkin, N. More and K. Roche [35] observed, in metallic multilayers of type {itinerant ferromagnetic metal (Co, Fe, etc.)/non-magnetic metal} $_{xN}$ , an interaction between the ferromagnetic layers which oscillated as a function of the thickness of the non-magnetic intercalating layer, with a period of nanometric order. The effect was then studied in a simpler three-layer structure, shown schematically in Fig. 14.15. The interaction energy per unit area between the two ferromagnetic layers  $F_1$  and  $F_2$  is given by

$$\varepsilon_{\text{int}} = -J\mathbf{m}_1 \cdot \mathbf{m}_2 = -J \cos(\theta_{12}) . \quad (14.29)$$

The unit vectors  $\mathbf{m}_1$  and  $\mathbf{m}_2$  represent the orientation of the magnetisations of layers  $F_1$  and  $F_2$ , assumed uniformly magnetised.  $\theta_{12}$  is thus the angle between  $\mathbf{m}_1$  and  $\mathbf{m}_2$ .  $J$  is the interlayer interaction constant. Positive and negative values of  $J$  favour parallel alignment (ferromagnetic interaction) or antiparallel alignment (antiferromagnetic interaction) of  $\mathbf{m}_1$  and  $\mathbf{m}_2$ , respectively. Figure 14.16 shows experimental measurements of  $J$  on a three-layer system [36].



**Fig. 14.16.** Experimental measurement of the oscillating interaction between ferromagnetic layers through a metallic layer [36]. The *continuous curve* represents a fit with the theory described in the text

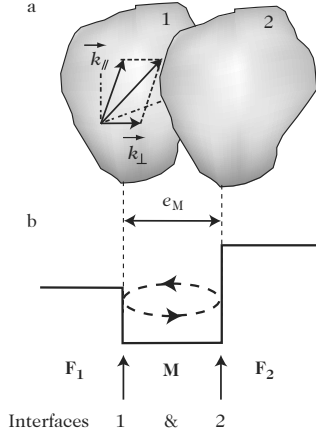
**Note.** The expression (14.29) describes an energy per unit area. To compare it with the other magnetic energies described in Sect. 14.1.3, which are expressed per unit volume, one must calculate the total energy of the system over its whole volume.

At this time, very few observations of quantum size effects had been made on metallic layers, and none concerned the electron spin. Coming just after the discovery of giant magnetoresistance in multilayers of the same type [1, 2], these results immediately triggered intense interest, leading to the observation of many magnetic quantum effects arising from spin-dependent confinement of electrons in magnetic multilayers. A detailed treatment can be found, for example, in the review article by P. Bruno [37]. However, the phenomenon and its extension to other quantities can be understood from an outline of the calculation using a free electron model at zero temperature.

#### *Electron Confinement in a Metallic Layer*

To begin with, let us ignore the magnetic aspects of the problem. The experimental triple layer in Fig. 14.15 can be simplified to the system shown in Fig. 14.17a, which assumes that the layers  $F_1$  and  $F_2$  have infinite extent in the perpendicular direction, and that they are perfectly plane and parallel. The electrons which go from  $F_1$  to  $F_2$  through the intercalating metallic layer  $M$  must therefore get past a potential step at each interface  $F_1/M$  and  $M/F_2$ . This is the well known problem of a particle getting past a potential well or barrier, the formal analogue of the optical Fabry–Perot interferometer discussed above [4]. In a free electron model, the profile of Fig. 14.17b corresponds to the profile of the bottom of the conduction band.

Let  $r_1 = |r_1|e^{i\varphi_1}$  and  $r_2 = |r_2|e^{i\varphi_2}$  be the reflection coefficients of the wave function of an electron travelling in the intercalating layer  $M$ . Subscripts 1



**Fig. 14.17.** Simplified model for calculating spin-dependent electron confinement. **(a)** Perspective view of the interfaces 1 and 2 between the ferromagnetic layers  $F_1$  and  $F_2$  and the metallic layer M. **(b)** Type of potential energy profile felt by the electron as it crosses M

and 2 refer to the interfaces M/ $F_1$  and M/ $F_2$ , respectively. The coefficients  $|r_1|$  and  $|r_2|$  determine the amplitude of the reflected wave and the phases  $\varphi_1$  and  $\varphi_2$  represent the phase shift during reflection. As for a photon in the Fabry–Perot device, when the electron wave makes a return trip in the layer M, its phase shifts by

$$\Delta\varphi = 2k_{\perp}e_M + \varphi_1 + \varphi_2, \quad (14.30)$$

where  $e_M$  is the thickness of the layer M and  $k_{\perp}$  is the perpendicular component of the wave vector  $\mathbf{k} = \mathbf{k}_{\parallel} + \mathbf{k}_{\perp}$  associated with the electronic state (see Fig. 14.17a).

Constructive and destructive interference [ $\Delta\varphi = 2n\pi$  or  $2(n+1)\pi$ , respectively] lead to an increase (resonance) or decrease, respectively, in the electron density of states in the layer M. Let  $\Delta n(\varepsilon, e_M)$  be the shift in the density of states with respect to the density of states without confinement, for electron energy  $\varepsilon$ . Quantum interference is manifested by a change  $\Delta E$  in the total energy  $E$  of the three-layer system, where

$$\Delta E(e_M) = \int_{-\infty}^{\varepsilon_F} (\varepsilon - \varepsilon_F) \Delta n(\varepsilon, e_M) d\varepsilon. \quad (14.31)$$

The sum up to the Fermi level  $\varepsilon_F$  corresponds to a sum over all electrons in the layer M. It is equivalent to going into the reciprocal space and summing over all wave vectors  $\mathbf{k}$  corresponding to occupied electronic states. As the phase shift  $\Delta\varphi$  in (14.30) and  $\Delta n(\varepsilon, e_M)$  depend only on  $k_{\perp}$ , the sum over  $k_{\parallel}$  is independent of the sum over  $k_{\perp}$ , and summing over all  $\mathbf{k}$  amounts to summing over all possible  $k_{\parallel}$  the solutions of one-dimensional problems with  $k = k_{\perp}$ . Moreover, in the (rather general) case of low confinement ( $r_1, r_2 \ll 1$ ),  $\Delta n(\varepsilon, e_M)$  varies as  $\cos(\Delta\varphi)$ . It is then a straightforward matter to show that [37]



$$\Delta E(e_M) \approx -\frac{1}{2\pi^3} \text{Im} \left[ \int dk_{\parallel}^2 \int_{-\infty}^{\varepsilon_F} d\varepsilon |r_1| |r_2| e^{i\Delta\varphi} \right]. \quad (14.32)$$

The integration over  $k_{\perp}$  has been transformed into an integration over  $\varepsilon$  by using the usual expression for the energy of an electronic state, viz.,

$$\varepsilon = \frac{\hbar^2}{2m} \left( k_{\parallel}^2 + k_{\perp}^2 \right).$$

In an obvious way, (14.32) is a sum of sinusoidal functions with continuously variable period. It cannot therefore produce any specific oscillation as a function of  $e_M$  unless there is a singularity in the integrand, capable of extracting a period from the continuum.

Consider first the integration over  $\varepsilon$ . As the Fermi distribution has a major singularity at  $\varepsilon = \varepsilon_F$ , we may keep only the contributions from electronic states at the Fermi surface. When we subsequently integrate over  $k_{\parallel}$ , a periodic oscillation will emerge for each vector  $k_{\parallel}$  corresponding to a vector  $2k_{\perp}$  that is stationary on the Fermi surface, i.e., a vector  $2k_{\perp}$  whose value barely depends on a variation of  $k_{\parallel}$  about a central value: the weight of the corresponding sinusoidal function is then increased in the integral. The oscillation period  $\Lambda_0$  is thus given by  $\Lambda_0 = 2\pi/2k_{\perp}$ , and it is an intrinsic characteristic of M.

A practical way of identifying these stationary vectors is to seek the vectors  $2k_{\perp}$  connecting two points on the Fermi surface whose Fermi velocities (perpendicular to this surface) are antiparallel. The amplitude of the oscillation increases with the radius of curvature of the Fermi surface at these points.

Finally, the confinement parameters, i.e., the electronic band mismatches between the metal M and the ferromagnetic metals  $F_1$  and  $F_2$ , affect the amplitude and phase of the oscillation of  $\Delta E$  through the reflection coefficients  $r_1$  and  $r_2$ .

The above argument was based upon a continuous medium approximation, valid for a free electron gas. The Fermi surface is then a sphere and the only stationary vector  $2k_{\perp}$  is the one measuring the diameter of the sphere for  $k_{\parallel} = 0$ . The period of oscillation is given by  $\Lambda_0 = \lambda_F/2$ , where  $\lambda_F$  is the Fermi wavelength of the electrons.

In the more realistic case of a crystal, the argument only remains valid if the interatomic distance is infinitesimal compared with  $\lambda_F$ . Now, in metals,  $\lambda_F$  is generally less than the nanometer. Introducing a crystal lattice into the model leads to two major modifications:

- The thickness  $e_M$  must be a whole number of atomic planes in the relevant crystal direction. The actual period of oscillation  $\Lambda$  thus results from regular sampling of the oscillation with period  $\Lambda_0$  given by the continuum model, with interval  $d$  equal to the distance between atomic planes. When  $d$  is greater than  $\Lambda_0/2$ , one speaks of aliasing, a phenomenon that is well known to signal processing specialists because it limits the spectral range of a digitally recorded signal, e.g., in an audio CD. The period is then given by

$$\frac{2\pi}{A} = q_{\perp} = \left| 2k_{\perp} - n \frac{2\pi}{d} \right|, \quad (14.33)$$

where  $n = 0$  or  $1$  depending on the initial value of  $2k_{\perp}$ . In fact,  $2\pi/d$  is the smallest vector of the reciprocal lattice in the direction perpendicular to the film.

- The integral over  $k_{\parallel}$  must be replaced by a discrete sum over all crystal sites on the interfaces. More generally, a full consideration of the crystal lattice leads to the result that two vectors  $k$  differing by a vector of the reciprocal lattice represent the same state. The stationarity condition on the vector  $k_{\perp}$  must therefore be sought on the periodic zone diagram, possibly between two different parts of the Fermi surface. This modification removes the limitation on the number of possible oscillations with different periods, which may therefore be greater than one [38, 39]. An example is discussed below.

#### *Oscillating Interaction Between Magnetic Layers Via a Non-Magnetic Layer*

In the magnetic trilayer of Fig. 14.16, the presence of an interaction between  $F_1$  and  $F_2$  means that the total energy of the system also depends on the respective orientations of the vectors  $m_1$  and  $m_2$ . This phenomenon is easily understood if we consider the specific feature of itinerant ferromagnetic metals that their conduction band is spin polarised (see Fig. 14.3). As a consequence, in an  $F_1/M/F_2$  trilayer system, the potential steps at the interfaces and the reflection factors  $r_1$  and  $r_2$  depend on the orientation of the electron spin with respect to the magnetisation of the ferromagnetic layer.

The energy  $\Delta E$  due to quantum interference is thus different depending on whether the layers  $F_1$  and  $F_2$  have parallel or antiparallel magnetisation. The problem is illustrated schematically in Fig. 14.18. For each interface, we define the spin asymmetry of the reflection coefficient:

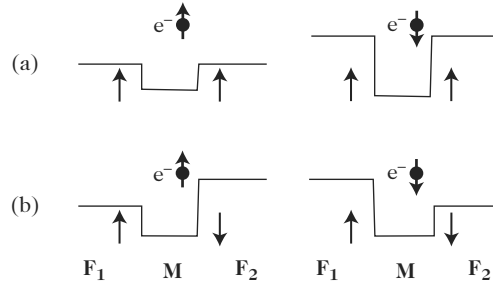
$$\Delta r_1 = \frac{r_1^{\uparrow} - r_1^{\downarrow}}{2}, \quad \Delta r_2 = \frac{r_2^{\uparrow} - r_2^{\downarrow}}{2}. \quad (14.34)$$

From (14.32), it is then easy to show that the interaction constant between layers  $F_1$  and  $F_2$  is given by

$$J(\epsilon_M) = \frac{E_{AF} - E_F}{2} \approx \frac{1}{\pi^3} \text{Im} \left[ \int dk_{\parallel}^2 \int_{-\infty}^{\epsilon_F} d\epsilon \Delta r_1 \Delta r_2 e^{2ik_{\perp} \epsilon_M} \right], \quad (14.35)$$

where  $E_{AF}$  and  $E_F$  denote the total energy for antiferromagnetic and ferromagnetic configurations of  $F_1$  and  $F_2$ , respectively.

We have the same kind of integral as in the non-magnetic case, and the selection rules for the oscillation periods and their intrinsic intensity, related to the Fermi surface of the metal  $M$ , will thus be the same. Only the amplitude and phase of the oscillation depend on the ferromagnetic metals via the confinement characteristics of the electrons.



**Fig. 14.18.** Spin-dependent energy profile for an electron  $e^-$  as it transits between ferromagnetic layers  $F_1$  and  $F_2$  via the non-magnetic layer  $M$ . Magnetisations of layers  $F_1$  and  $F_2$  oriented (a) parallel (ferromagnetic configuration) and (b) antiparallel (antiferromagnetic configuration)

In the general case, we obtain the following expression for  $J$ :

$$J(e_M) = \sum_{\alpha} K_{\alpha} \frac{1}{e_M^2} \cos(q_{\perp\alpha} + \varphi_{\alpha}), \quad (14.36)$$

where the sum is taken over all possible stationary vectors, indexed by  $\alpha$ . The search for quantum size effects in multilayers, whether magnetic or otherwise, has been intense and has led to some debate. This is not the place to go into a full discussion and we shall merely describe some illustrative examples.

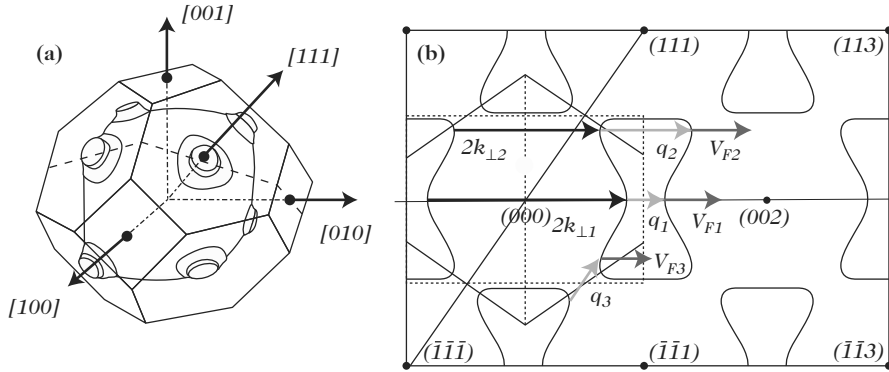
Quite generally, intercalating layers  $M$  of noble metals (Cu, Ag, Au) have provided the most accurate observations. We shall therefore discuss this case in more detail. The Fermi surface of these metals is like that of Cu, shown schematically in Fig. 14.19. It only deviates from the sphere by a kind of neck formation in the (111) directions. It is therefore rather easy to identify the stationary vectors  $k_{\perp}$  which give the interaction oscillations, at least in the two commonest cases [38, 39].

- If the interface is perpendicular to the [001] direction, the vector measuring the diameter of the surface for  $k_{\parallel} = 0$  is denoted by  $k_{\perp 1}$  in Fig. 14.19b. The oscillation period is given by the vector

$$q_{\perp 1} = \frac{2\pi}{\Lambda} = \left| 2k_{\perp 1} - \frac{2\pi}{d} \right|,$$

which differs from  $2k_{\perp 1}$  by a vector of the reciprocal lattice. There is a second stationary vector, denoted by  $k_{\perp 2}$ , for a nonzero value of  $k_{\parallel}$ . Here too, the period is obtained after subtracting the vector from the reciprocal lattice.

- If the interface is perpendicular to the [111] direction, there is only one stationary vector  $k_{\perp}$ , denoted by  $k_{\perp 3}$ , around the necks in the (111) direction. The oscillation period in this case is given directly by  $\Lambda = \pi/k_{\perp 3}$ .



**Fig. 14.19.** Determining the oscillation periods in a copper layer. (a) Perspective view of the Fermi surface of copper. (b) Cross-sectional view through the plane  $\{[001], [111]\}$ . When  $[001]$  is perpendicular to the layer M, the vectors  $2k_{\perp 1}$  and  $2k_{\perp 2}$  are the two stationary vectors giving the quantum size effects. The corresponding periods are given by the vectors  $q_1$  and  $q_2$ , which differ by the smallest vector of the reciprocal lattice in the direction  $[001]$ . When  $[111]$  is perpendicular to M, there is only one stationary vector, viz.,  $2k_{\perp 3} = q_3$ . The Fermi velocities of the electrons corresponding to the three oscillations are denoted by  $v_{F(1,2,3)}$

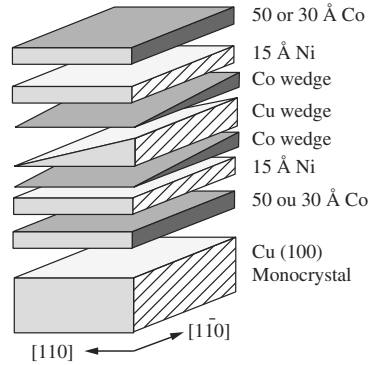
**Table 14.2.** Theoretical values of the periods of the interlayer oscillating interaction through an intercalating layer M made from a noble metal. Values are given in numbers of atomic planes (AP)

Orientation		Cu	Ag	Au
[001]	$d_{001}$ [Å]	1.80	2.03	2.03
	$A_1$ [AP]	5.88	5.58	8.60
	$A_2$ [AP]	2.56	2.36	2.51
[111]	$d_{111}$ [Å]	2.08	2.35	2.35
	$A_3$ [AP]	4.5	5.94	4.83

Table 14.2 gives the values of these periods for three noble metals.

The most complete characterisation of the oscillating interaction has been achieved for the system  $\text{Fe}/(\text{Au,Ag})(001)/\text{Fe}(100)$ , where the perfect growth of the noble metal on a whisker (monocrystal) of  $\text{Fe}(100)$  with atomically flat surface has provided observations of biperiodic oscillations up to thicknesses  $e_M$  greater than 50 atomic planes [40]. In this case, measured parameters have also been compared to very high accuracy with theoretical predictions [41]. For all the other systems, the oscillations are damped to differing degrees by interface defects (see Appendix A).

Spin-dependent oscillations of the electron density of states were directly observed for the first time on the  $\text{Cu}/\text{Co}(001)$  system by photoemission



**Fig. 14.20.** Sample with wedge-shaped layers used to measure quantum size effects in ultrathin layers [47]

measurements during growth [42, 43]. One interesting consequence is the existence of an oscillation in the spin polarisation of the electrons in the metal M.

When the metal M is replaced by an alloy, a continuous modification of the Fermi surface can be observed. For example, Cu and Ni are neighbours in the periodic table, with Ni having one electron fewer than Cu. By increasing the concentration of Ni in a  $\text{Cu}_{1-x}\text{Ni}_x$  alloy, the average number of electrons per atom is changed. In a rigid electron band model, this shows up through a contraction of the Fermi surface. The effect is expected to be more pronounced near the necks and a rapid increase in the period of oscillation along the [111] direction has indeed been observed when the Ni concentration is increased [44, 45].

Cr is a special case. In the [001] direction, two oscillations are observed, one with long period (10–12 atomic planes) and the other with very short period (roughly 2 atomic planes) [46]. The latter oscillation has a very high amplitude. In fact, at the stationary vector  $k_{\perp}$  underlying the oscillation, the Fermi surface has an almost infinite radius of curvature in the two directions parallel to the interface. This feature is indeed responsible for the antiferromagnetism of Cr in the bulk solid state.

In order to obtain very accurate thicknesses, the above experiments were carried out using samples in which the layers of metal M were wedge-shaped, as shown in Fig. 14.14. More complex samples were then composed by superposing several wedge-shaped layers, as shown in Fig. 14.20. These samples revealed quantum size effects in the ferromagnetic layers themselves [47, 48], which we have assumed to be infinite up to now, but also in the metallic coating layers deposited on the trilayers to protect them [94]. These effects modulate the classical oscillating exchange interaction.

Theoretical predictions concerning the dependence of the phase of the oscillation on the ferromagnetic metal were also confirmed [49].

Note also that the interlayer oscillating coupling was first observed in 1987 [50] for rare earth multilayers  $[\text{Gd}/\text{Y}]_{xN}$ . This can be interpreted in terms

of the Ruderman–Kittel–Kasuya–Yosida interaction, rather than through electron confinement effects, but the relationship with the Fermi surface remains the same [38, 39].

The theory of the interlayer oscillating interaction extends to insulating intercalating layers [37] by introducing complex wave vectors for the electron states in the barrier. The factor  $\exp(ikD)$  then includes a strong exponential damping component which corresponds effectively to an electron wave transmitted by the tunnel effect. The interaction thus has a very short range in this case.

#### *Other Examples of Spin-Dependent Electron Confinement Effects*

The above calculation is only a specific example of a quantum size effect, although rather spectacular, since it reveals a novel effect of considerable amplitude. More generally, electron confinement in a magnetic layer can induce oscillating behaviour in a great many quantities by modifying the electron density of states. However, the confinement effect is generally much more complex.

For example, strong oscillations of the Kerr magneto-optical effect have been observed both with respect to the thickness of a ferromagnetic layer [51] and with respect to the thickness of a metal layer coating the ferromagnetic layer [52]. But the measured periods have no connection with the Fermi surface, as can easily be understood. The Kerr effect corresponds to the rotation of the light polarisation during reflection by a magnetic material. This rotation is due to electron radiative transitions from occupied states to free states in the presence of the exchange shift of electron bands (see Sect. 14.1.1) and the spin–orbit interaction (see Sect. 14.1.2). Quantum size effects can modify the initial or final densities of states of the transition, and hence its intensity. But it is the presence of a vertical transition at the photon energy which selects an oscillation period of the continuum, rather than the integral over a singularity. It is therefore much more difficult to calculate oscillation periods [53].

Likewise, it has been observed that the interface magnetic anisotropy of an ultrathin layer of Co can oscillate with the thickness of a Cu coating layer, although it has not been possible to identify accurately which states are involved.

#### **14.1.5 Magnetisation Dynamics in Magnetic Nanostructures**

All the principles of magnetisation dynamics can be exposed by restricting to a magnetic nanostructure, generally defined as a magnetic element that is small enough to ensure that at any instant the magnetisation is kept uniform by the exchange interaction. This condition is respected in nanoparticles with size less than the exchange length  $\lambda_{\text{exch}}$  (see Sect. 14.1.3). We shall also assume here that the magnitude  $M_S$  of the magnetisation remains constant:  $\mathbf{M} = M_S \mathbf{m}$ . One then speaks of macrospin to describe the particle and coherent process

for its magnetisation dynamics. Bigger elements but with strictly ellipsoidal shape can also be considered as macrospins in certain situations, but this is not a very common situation in the experimental context. In other situations, the effect of the demagnetising field (see p. 512) generally makes the magnetisation non-uniform and its temporal and spatial fluctuations can strongly influence the magnetisation reversal process.

In this section, we shall consider only macrospin dynamics, although we shall mention the influence of a larger size and non-ellipsoidal shape in several important cases of real magnetic elements. Generally speaking, the dynamics of larger elements can be calculated by a numerical finite element method, applying the basic principles to each cell in the calculation and coupling the cells together by the various magnetic interactions, e.g., exchange, dipole, interlayer, etc. [54].

### Precession: Basic Dynamics

The dynamics of the magnetisation of a macrospin is described by the fundamental Landau–Lifshitz–Gilbert (LLG) equation [54–56], which applies in the vast majority of cases where the magnitude of the magnetisation remains constant, as we have assumed since the beginning of the chapter:

$$\frac{d\mathbf{M}(t)}{dt} = -\gamma[\mathbf{M}(t) \times \mu_0 \mathbf{H}_{\text{eff}}] + \frac{\alpha}{M_S} \left[ \mathbf{M}(t) \times \frac{d\mathbf{M}(t)}{dt} \right], \quad (14.37)$$

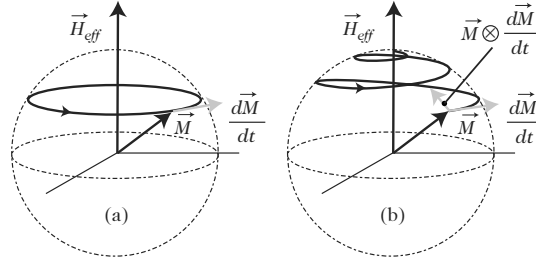
where  $\mathbf{H}_{\text{eff}}$  is the effective magnetic field felt by the magnetisation at each instant of time. The first term on the right-hand side of this equation describes the couple exerted by the magnetic field on the magnetic moment of the macrospin [see (14.2)]. It arises from quantum mechanics and indeed the gyromagnetic factor  $\gamma$  which appears in the equation is given by  $\gamma = g\mu_B/\hbar$ , involving Planck's constant. The second term on the right-hand side expresses the energy dissipated during the motion, described in the form of a fluid damping characterised by the damping factor  $\alpha$ .

From (14.37), it can be shown that energy is lost at the rate [54]

$$\frac{dE}{dt} = -\frac{\alpha}{\gamma M_S} \left( \frac{d\mathbf{M}}{dt} \right)^2. \quad (14.38)$$

This rate therefore increases as the magnetisation rotates more quickly. Equation (14.37), which is clear from a physical point of view, is difficult to handle because the derivative of  $M$  appears on both sides of the equals sign. The following mathematically clearer form is thus preferred:

$$(1 + \alpha^2) \frac{d\mathbf{M}}{dt} = -\gamma(\mathbf{M} \times \mu_0 \mathbf{H}_{\text{eff}}) + \gamma \frac{\alpha}{M_S} (\mathbf{M} \times \mathbf{M} \times \mu_0 \mathbf{H}_{\text{eff}}). \quad (14.39)$$



**Fig. 14.21.** Trajectory of the magnetisation of a macrospin when  $H_{\text{eff}}$  is constant. (a) Precession without damping. (b) Relaxation towards the minimal energy orientation when energy is dissipated

#### *Trajectory for an Isotropic Magnetic Nanostructure*

Consider first the case of an isotropic magnetic nanostructure.  $\mathbf{H}_{\text{eff}}$  is then equal to the magnetic field  $\mathbf{H}_{\text{ext}}$  created by external sources. Since the magnitude  $M_S$  is constant, the end of the vector  $\mathbf{M}$  describes trajectory on a sphere of radius  $M_S$ , and  $d\mathbf{M}(t)/dt$  is always perpendicular to  $\mathbf{M}$ . If there is no dissipation, i.e.,  $\alpha = 0$ , then  $d\mathbf{M}(t)/dt$  also remains perpendicular to the plane containing  $\mathbf{M}$  and  $\mathbf{H}_{\text{ext}}$ , whereupon the magnetisation precesses around the field direction  $\mathbf{H}_{\text{ext}}$  without energy loss, as illustrated in Fig. 14.21a. The angular speed  $\omega_0$  is given by

$$\omega_0 = \gamma\mu_0 H_{\text{eff}}, \quad (14.40)$$

which corresponds to a precession frequency of

$$f = \frac{\omega_0}{2\pi} = \frac{\gamma}{2\pi} \mu_0 H_{\text{eff}}.$$

The factor  $\gamma/2\pi$  is of the order of 28 MHz/mT for a free electron, and the same value is approximately valid for itinerant ferromagnetic metals where the orbital moment is quenched. Considering that a field of the order of 100 mT is fairly easily obtained in magnetic devices, the corresponding precession frequency  $f = 2.8$  GHz is a good order-of-magnitude estimate of the rate at which the magnetisation can be reversed.

To a first approximation ( $\alpha \ll 1$ ), the dissipative term

$$\mathbf{M}(t) \times \frac{d\mathbf{M}(t)}{dt}$$

is oriented towards  $\mathbf{H}_{\text{ext}}$  and transforms the motion into a damped precession (see Fig. 14.21b) which gradually drags the magnetisation into the direction of  $\mathbf{H}_{\text{ext}}$ , i.e., into the direction of minimal energy. The factor  $\alpha$  thus characterises the rate at which the magnetisation returns to equilibrium.

It is not easy to determine the origin of the damping effect, which is hidden, sometimes rather crudely, behind the single phenomenological parameter  $\alpha$ :



- So-called intrinsic dissipation refers to energy transfer towards the various heat reservoirs, e.g., electrons, phonons, magnons, etc., during physical processes. For example, due to magnetostriction, the shape of a nanostructure fluctuates with the magnetisation direction. When the magnetisation precesses, it thus generates phonons, even in the material supporting the nanostructure, thereby raising its temperature. The effect must generally be decomposed into several complex processes which may be facilitated by extrinsic defects such as impurity atoms and bulk or interface crystal defects.
- There may also be an inhomogeneous contribution to  $\alpha$  which does not directly correspond to energy dissipation. For example, in an ensemble of magnetic particles, a distribution of properties leads to a distribution of precession rates, and this gradually destroys the uniformity in the direction of magnetisation of the particles. If the mean moment of the ensemble is measured, one observes a more rapid apparent return to equilibrium than what is actually undergone by each particle and the factor  $\alpha$  measured over the ensemble is greater than the corresponding factor for an individual particle. This argument can be extended to a large-scale non-ellipsoidal magnetic element, which is not therefore uniformly magnetised.

*General Case of an Anisotropic Magnetic Nanostructure*

Real magnetic nanostructures are never perfectly isotropic. In the general case,  $\mathbf{H}_{\text{eff}}$  depends on the magnetic energy per unit volume  $\varepsilon_{\text{mag}}$  via the expression

$$\mathbf{H}_{\text{eff}} = -\frac{1}{\mu_0 M_S} \frac{\delta \varepsilon_{\text{mag}}}{\delta \mathbf{m}} = \mathbf{H}_{\text{ext}} + \mathbf{H}_{\text{exch}} + \mathbf{H}_{\text{D}} + \mathbf{H}_{\text{A}} . \quad (14.41)$$

The magnetic fields appearing on the right-hand side correspond to the various energy sources discussed in Sect. 14.1.3.  $\mathbf{H}_{\text{ext}}$  is simply the external field (Zeeman energy).  $\mathbf{H}_{\text{exch}}$  arises from the exchange energy and is given by

$$\mathbf{H}_{\text{exch}} = \frac{2A}{\mu_0 M_S} \nabla^2 \mathbf{m} .$$

In a nanostructure, the magnetisation is uniform and hence  $\mathbf{H}_{\text{exch}} = 0$ .  $\mathbf{H}_{\text{D}}$  is the usual demagnetising field (see p. 512), given by

$$\mathbf{H}_{\text{D}} = -M_S N \mathbf{m} , \quad (14.42)$$

where  $N$  is the tensorial demagnetisation factor.

$\mathbf{H}_{\text{A}}$  represents the anisotropy energy and can thus assume complex forms (see Sect. 14.1.3). It will be general enough for our purposes to consider a uniaxial anisotropy with easy axis along the  $Ox$  direction, where the anisotropy energy is given by  $\varepsilon_{\text{A}} = -K m_X^2$  and the anisotropy field by

$$\mathbf{H}_{\text{A}} = \frac{2K}{\mu_0 M_S} m_X \mathbf{x} , \quad (14.43)$$

where  $\mathbf{x}$  is the unit vector in the  $Ox$  direction. As in the case of ultrathin films, the anisotropy constant  $K$  represents an average over the nanostructure. In a rather inexact way, the maximal amplitude of  $\mathbf{H}_A$ , viz.,

$$H_A = \frac{2K}{\mu_0 M_S},$$

is often referred to as the anisotropy field.

Obviously, the effective field varies, generally rather significantly, with the motion of the magnetisation, and the trajectories are complex. The important example of a wafer-shaped nanostructure is described in Appendix B, where it is shown how these precession effects can be used to optimise magnetisation reversal.

Finally, in a slightly bigger sample, where the magnetisation is not strictly uniform, the effective field can also vary spatially, e.g., under the influence of the demagnetising field or the exchange field, which is no longer zero. As mentioned above, these fluctuations contribute to the experimentally measured damping coefficient  $\alpha$ .

### Quasi-Static Reversal at Zero Temperature

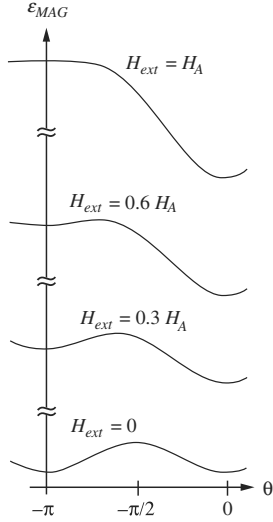
We have just seen that precessional motion corresponds to frequency of GHz order. In most experimental situations, the external field  $\mathbf{H}_{\text{ext}}$  varies much more slowly than this. Under the effect of damping, the magnetisation  $\mathbf{M}$  has time to relax at each instant towards the nearest magnetic energy minimum. This minimum is not generally a stable state, i.e., a global energy minimum, but rather a metastable state separated from the stable state by an energy barrier. At zero temperature, when there is no heat energy, the magnetisation remains blocked in this metastable state until the barrier disappears under the effect of the field  $H_{\text{ext}}$ . This behaviour is simply illustrated in the case of a spherical nanostructure ( $N$  isotropic) with uniaxial magnetic anisotropy having easy axis in the  $Ox$  direction. In the initial state, the magnetisation lies along the direction  $-\mathbf{x}$ .

#### $\mathbf{H}_{\text{ext}}$ Applied Along the Easy Axis

Let us assume to begin with that the external field  $\mathbf{H}_{\text{ext}}$  lies in the  $+\mathbf{x}$  direction. The magnetic energy per unit volume is then given by

$$\varepsilon_{\text{mag}} = K \sin^2 \theta - \mu_0 M_S H_{\text{ext}} \cos \theta, \quad (14.44)$$

where  $\theta$  is the angle between the magnetisation and  $+\mathbf{x}$ . Figure 14.22 shows how the energy profile  $\varepsilon_{\text{mag}}$  changes when  $H_{\text{ext}}$  increases. Note the metastable energy minimum corresponding to the initial orientation  $\theta = \pi$ , separated from the stable state  $\theta = 0$  by an energy barrier which disappears when  $H_{\text{ext}}$  reaches the value of the anisotropy field  $H_A$ . The magnetisation then flips



**Fig. 14.22.** Energy profiles as a function of the direction of magnetisation for a uniaxial anisotropic nanostructure in an external field  $H_{\text{ext}}$  lying along  $\theta = 0$

from the  $-\mathbf{x}$  direction into the  $+\mathbf{x}$  direction (see the curve labelled  $0^\circ$  in Fig. 14.23A).

The height  $E_B$  of the barrier separating the stable and metastable orientations is given by

$$E_B = VK \left(1 - \frac{H}{H_A}\right)^2, \quad (14.45)$$

where the volume  $V$  of the sample arises because this is the total energy.

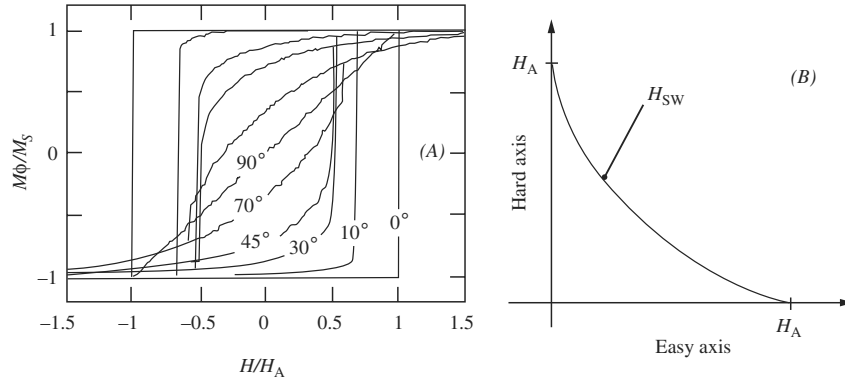
#### *$H_{\text{ext}}$ Applied in an Arbitrary Direction*

The system has axial symmetry about the easy magnetisation axis. Without loss of generality, we may therefore restrict to the plane containing this axis and  $\mathbf{H}_{\text{ext}}$ . For an arbitrary angle  $\phi$  between  $\mathbf{H}_{\text{ext}}$  and the easy axis, reversal will occur for a field  $H_{\text{SW}}$ , the Stoner–Wohlfarth field, given by

$$\frac{H_{\text{SW}}}{H_A} = \frac{1}{\left(\sin^{2/3} \phi + \cos^{2/3} \phi\right)^{3/2}}. \quad (14.46)$$

Figure 14.23A shows a series of hysteresis cycles in the magnetisation for several values of the angle  $\phi$ . In contrast to the case in which the applied field lies along the easy axis ( $\phi = 0$ ), for an arbitrary orientation of  $\mathbf{H}_{\text{ext}}$ , reversal is preceded and followed by a phase in which the magnetisation rotates.

Equation (14.46) defines the Stoner–Wohlfarth astroid, named after the two scientists who first studied this behaviour [57]. One also speaks of the coherent Stoner–Wohlfarth reversal mode. The shape of this astroid is illustrated in Fig. 14.23B. A. Thiaville recently generalised this study to the case



**Fig. 14.23.** (A) Hysteresis cycle for the component of magnetisation along the external field  $\mathbf{H}_{\text{ext}}$  as a function of the relative amplitude of  $H_{\text{ext}}$  with respect to the anisotropy field  $H_A$ , for several values of the angle between  $\mathbf{H}_{\text{ext}}$  and the easy axis (in degrees). (B) Stoner–Wohlfarth astroid giving the field  $H_{\text{SW}}$  for irreversible reversal of the magnetisation for a field  $\mathbf{H}_{\text{ext}}$  with arbitrary direction. The curve is only drawn for positive  $H_x$  and  $H_y$ , but the rest is easily obtained by symmetry

of arbitrary anisotropy [58] in order to explain new experimental observations on magnetic nanoparticles [59, 60]. The expression for the barrier energy has also been generalised [61, 62].

When the particle size exceeds  $\Lambda_{\text{exch}}$ , the exchange interaction is no longer a priori strong enough to keep the magnetisation uniform during a reversal process. Complex non-uniform reversal modes then occur, depending on the size and geometry of each particle. These modes generally reduce the reversal field. For an element with uniaxial magnetic anisotropy such as we have been considering up to now, the effect increases as the direction of the external magnetic field approaches the easy axis, and this distorts the astroid in Fig. 14.23B. The effect has been studied quantitatively on nanowires, for example [16].

Finally, a recent experiment has led to a unification of the precessional and coherent reversal effects of Stoner–Wohlfarth type. Subjecting Co nanoparticles to a radio frequency pulse (several GHz), Thirion et al. [63] observed reversals at much smaller fields than the Stoner–Wohlfarth field. The experiment shows that the astroid is distorted by resonances which in fact correspond to the excitation of favoured precessional motions of the particle magnetisation at the bottom of the metastable energy well. If the input radio frequency power is great enough, the precession allows the particle to overcome the energy barrier.

### Thermally Activated Quasi-Static Reversal

At nonzero temperatures, the presence of thermal activation can help the magnetisation to overcome a nonzero energy barrier and thereby reach a stable state. This phenomenon was predicted by L. Néel as early as 1949 [64], and further studied by W.F. Brown [65], who sought to account for thermal excitations theoretically.

In the Néel–Brown model, the probability of the magnetisation of a magnetic nanoparticle not being reversed by thermal activation after time  $t$  is given by the exponential law

$$P = e^{-t/\tau} . \quad (14.47)$$

The time  $\tau$  characterising thermal stability is in turn given by an Arrhenius law:

$$\tau = \tau_0 \exp \frac{E_B(H, T)}{k_B T} . \quad (14.48)$$

The barrier height  $E_B$  depends on the applied field, of course, but also on the temperature via the temperature dependence of the anisotropy parameters, for example.  $k_B$  is the Boltzmann factor ( $k_B = 1.380\,54 \times 10^{-23}$  J/K), whence  $k_B T$  represents the thermal activation energy.  $\tau_0$  is an intrinsic characteristic reversal time depending on the properties of the nanostructures. One also speaks of the attempt frequency  $f_0 = 1/\tau_0$ . In the Brown model, thermal activation is taken into account by means of a random magnetic field [65, 66], which excites the resonant precession frequencies of the magnetisation around the direction corresponding to minimum energy.  $\tau_0$  thus depends on the form of the anisotropy energy and also the damping factor  $\alpha$  [67]. It is very difficult to calculate  $\tau_0$ , or to measure it directly, but in typical systems, it has a value of nanosecond order. For example, the first direct observation of the above laws made in 1997 by W. Wernsdorfer et al. on ellipsoidal Co particles with sizes  $\sim 25$  nm gives  $\tau_0 \sim 3 \times 10^{-9}$  s [67, 68]. A more recent observation on Co clusters with sizes  $\sim 3$  nm [60] gives  $\tau_0 \sim 10^{-10}$  s.

The above expressions have two important consequences for applications of nanomagnetism at room temperature:

- The first is that the magnetic storage of binary information is marred by an intrinsic error rate. Suppose for example that a large amount of binary data is stored in the magnetisations of an array of identical non-interacting particles, with volume  $V$  and anisotropy constant  $K$ . In zero field,  $E_B = KV$ . From the above equations, one calculates that after 10 years, i.e.,  $\sim 3 \times 10^8$  s, with  $KV = 54k_B T$ , the error rate ( $1 - P$ ) due to magnetisation reversal by thermal activation will be  $10^{-6}$ , that is, one error per 1 Mbit. One must go to  $KV = 68k_B T$  to achieve an error rate of  $10^{-12}$ , that is, about one error per 100 Gbytes, the capacity of a hard disk. In the quest for very high recording densities,  $KV$  must be reduced to this

type of value, whilst the dipole interaction makes the problem worse still. To maintain a negligible error rate, computer error correction codes are used. In a hard disk, each bit of information is also stored on a group of a hundred or so particles.

- The second consequence is that the external field  $\mathbf{H}_{\text{SW}}$  required to reverse the magnetisation of a particle decreases when its time of application  $\delta t$  is increased. For example, for a field applied along the easy magnetisation axis, the variation is easily obtained from (14.45) and (14.48) by inserting  $\tau = \delta t$ :

$$H_{\text{SW}}(T, \delta t) = H_A \left[ 1 - \sqrt{\frac{k_B T}{KV} \ln \left( \frac{\delta t}{\tau_0} \right)} \right]. \quad (14.49)$$

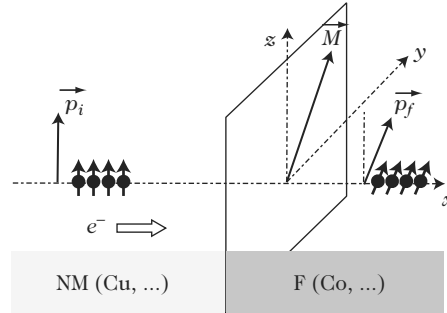
It must be remembered that this dependence is probabilistic. Note that, for  $\delta t = \tau_0$ ,  $H_{\text{SW}}$  reaches the value  $H_A$  of the anisotropy field, cancelling the energy barrier and thereby leading inevitably to reversal. In practice, the range of pulse widths  $\delta t$  around  $\tau_0$  and below, which are hard to reach, is not well known and has not yet seen much investigation [69].

Finally, at very low temperatures, where thermal activation is no longer able to help in overcoming the energy barrier, there should still be a low level of transition due to tunneling through the barrier. This phenomenon, known as the macroscopic quantum tunnel effect, has been the subject of active investigation over the last few years. A direct signature of this effect is a reversal rate that does not depend on the temperature when the sample is cooled below a certain critical temperature. Although this kind of behaviour has indeed been observed for nanoparticles with sizes around  $10^5$  atoms [70], unambiguous observation of a reversal by macroscopic tunnel effect has so far only been reported for high spin magnetic molecules. The reader is referred to the excellent review by W. Wernsdorfer [71], which also discusses thermal activation effects in some detail.

### Dynamics of Magnetisation Induced by a Current of Polarised Electrons

A last method for controlling magnetisation reversal in a magnetic nanostructure has recently been proposed by J. Slonczewski [72] and L. Berger [73]. The idea behind this method is shown schematically in Fig. 14.24.

In a ferromagnetic metal F such as a 3d transition metal, the spin of the conduction electrons interacts with the global magnetisation  $\mathbf{M}$  of the material. In ‘standard’ electron optics, if an electron with spin aligned along an orientation  $\mathbf{p}_i$  that is not parallel with  $\mathbf{M}$  is injected into F, this interaction generates a couple which, by damped precession, gradually lines up the electron spin with  $\mathbf{M}$ . As this happens, the spins abandon a transverse component of their angular momentum which is transferred to the magnetisation by the



**Fig. 14.24.** Schematic view of the model for spin angular momentum transfer. A flux of electrons with spin parallel to the unit vector  $\mathbf{p}_i$  is injected from a non-magnetic metal layer NM into a ferromagnetic layer F with magnetisation  $\mathbf{M}$ . Since the spins interact with  $\mathbf{M}$ , beyond a certain distance from the interface, the spin polarisation  $\mathbf{p}_f$  of the electrons will have rotated to line up with  $\mathbf{M}$ . By the principle of action and reaction, the electron spins will have transferred a transverse angular momentum to the magnetisation which will tend to line up  $\mathbf{M}$  with  $\mathbf{p}_i$

principle of action and reaction. This transverse angular momentum acts on  $\mathbf{M}$  to align it with  $\mathbf{p}_i$ , a process referred to as the spin transfer mechanism.

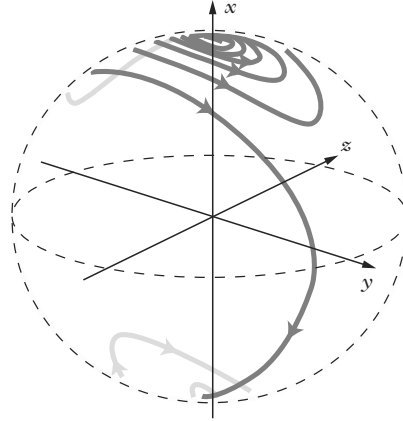
It is thus possible to act on a magnetisation using an electron current, if this current has spin polarisation, i.e., if there are more electrons with spins parallel to  $\mathbf{p}_i$  than antiparallel to it. A realistic calculation remains complex and is still the subject of some debate. It can nevertheless be shown that, in the Landau–Lifshitz–Gilbert equation applied to the magnetisation  $\mathbf{M}$ , the effect shows up through an extra effective field given by [74]

$$\mathbf{H}_{\text{inj}} = \chi [\mathbf{M} \times \mathbf{p}_i] + \beta \chi M_S \mathbf{p}_i . \quad (14.50)$$

The factor  $\chi$  measures the intensity of the effect. It is of course proportional to the number of injected electrons, and hence the current, and depends on the angle between  $\mathbf{p}_i$  and  $\mathbf{M}$  [72]. Moreover, it can be shown that the effect extends over a very short range ( $< 1$  nm) in the magnetic layer. It is thus an interface effect, averaged by the exchange interaction throughout the thickness  $e$  of a magnetic layer.  $\chi$  is thus proportional to  $1/e$ .

The factor  $\beta$  is estimated to be less than unity, and the first term in (14.50) thus dominates, even though the second may turn out to be significant in certain situations. If we integrate this first term in (14.50), we obtain the same kind of expression as the damping term proportional to  $\alpha$ . For this reason, it can be said that the spin transfer mechanism acts like a negative damping factor  $\alpha$ , i.e., one that brings energy to the precession rather than removing it.

To illustrate the magnetisation dynamics induced by this effect, consider the case of a nanoparticle with uniaxial anisotropy with easy axis lies along  $Ox$ , through which passes a spin-polarised current of intensity  $I$ . We neglect

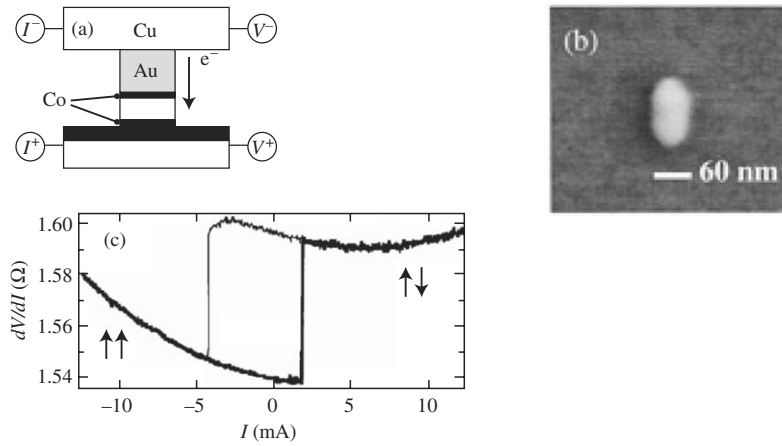


**Fig. 14.25.** Trajectory of the magnetisation during a reversal induced by the spin transfer mechanism. The trajectory has been calculated for a magnetic nanostructure in the form of a thin wafer, parallel to the  $xy$  plane and with easy magnetisation axis along the  $x$  axis. Initially, the magnetisation lies along  $+x$  with a slight misalignment and the spin polarisation of the injected electrons is oriented in the  $-x$  direction. Due to the dipole energy, the precession is flattened in the  $z$  direction

the term in  $\beta$  in (14.50). Note to begin with that, if the orientation  $\mathbf{p}_i$  of the spin polarisation is strictly antiparallel to the magnetisation  $\mathbf{M}$ , the couple exerted by the current is zero. To avoid this situation, we shall assume that, at  $t = 0$ , the magnetisation makes a small angle  $\delta\theta$  with the easy magnetisation axis along  $+x$ , whilst  $\mathbf{p}_i = -x$ . When the current is zero,  $\mathbf{M}$  precesses, losing energy until it is once again parallel to  $+x$ . When the current is increased from zero, two thresholds appear at  $I_1$  and  $I_2$ . These depend on all the parameters of the particle and the spin polarisation of the current. For  $I < I_1$ , the energy supplied is insufficient to prevent  $\mathbf{M}$  from relaxing towards  $+x$ . Above  $I_1$ , the precession of the magnetisation stabilises to a specific trajectory whose angle increases with the current. This trajectory corresponds to compensation of the energy dissipated during precession by the spin transfer mechanism. Finally, above  $I_2$ , the precession angle reaches  $180^\circ$  and the magnetisation flips over into the  $-x$  half-space before relaxing into the  $\mathbf{p}_i$  direction. The magnetisation of the particle has thus been reversed. This mechanism is illustrated in Fig. 14.25.

The trilayer structure shown in Fig. 14.26a has provided several demonstrations of this effect (see, for example, [75]). It involves spin-dependent electron transport effects which will be discussed in Sect. 14.2. We can nevertheless give a qualitative description referring to the current-dependent hysteresis cycle of magnetisation!hysteresis shown in Fig. 14.26b. The cycle shows how the magnetoresistance of the pillar varies with the injected current. Starting from the high positive current  $I^+$  where the value of the resistance indicates antiparallel alignment of the moments in the two layers (see Fig. 14.1 and



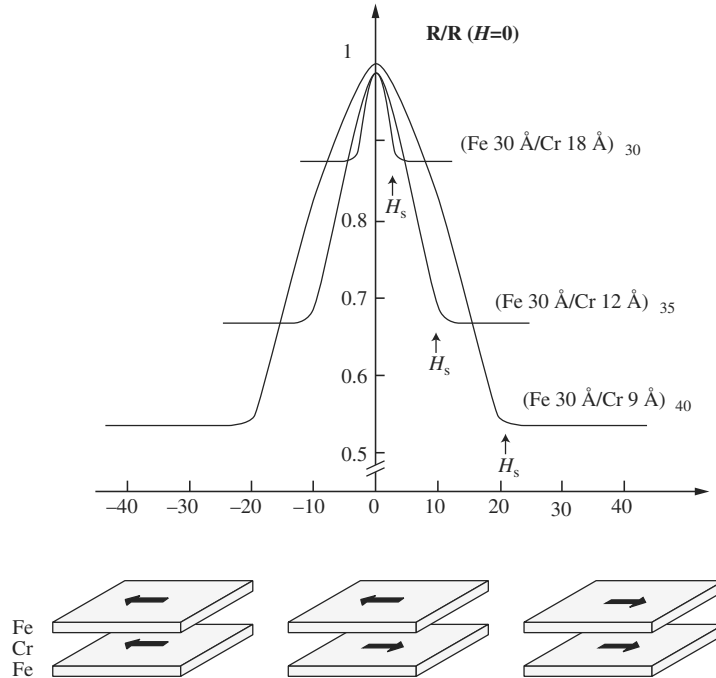


**Fig. 14.26.** Experimental observation of magnetisation reversal by transfer of spin angular momentum in a device etched in a Co/Cu/Co trilayer. **(a)** Schematic of measurement device. **(b)** Electron microscope image of the pillar during fabrication. *Lower:* Hysteresis cycle obtain by injecting a current  $I$  through the pillar [75]

Sect. 14.2), when electrons are injected from the base of the device (negative current  $I^-$ ), they acquire a spin polarisation as they cross the first rather thick ferromagnetic layer, and their interaction with the second rather thin layer (1–3 nm) favours alignment of its magnetisation parallel to the magnetisation of the thick layer. Over one cycle, one does indeed observe a transition towards a low resistance state for a negative value of the current, and this provides a direct illustration of the spin transfer mechanism discussed above. The explanation of the opposite transition for a positive current is more complex [76].

In a trilayer device of this kind, S.I. Kiselev et al. [77] have directly revealed the precession induced by injection of a constant current by measuring the microwave frequency noise across the nanomagnet terminals. Indeed, the resistance of the nanomagnet fluctuates as the magnetisation precesses.

Finally, this trilayer device is extremely useful in applications, since it can be used to write binary information by sending a current directly into the magnetoresistive device. Moreover, a deeper examination of the underlying equations of this mechanism shows that, for the same layer thickness, reversal occurs at constant current density when the lateral dimensions of the layer are reduced, a feature which clearly facilitates device miniaturisation.



**Fig. 14.27.** Resistance as a function of applied magnetic field measured in Fe/Cr multilayers [95]. When the applied field exceeds the field due to indirect antiferromagnetic exchange coupling between adjacent Fe layers, the resistance of the multilayer drops sharply, giving rise to the phenomenon known as giant magnetoresistance

## 14.2 Spin Electronics

### 14.2.1 Description

Spin electronics is a new branch of electronics, based not only on the charge, electron or hole, of carriers within semiconducting structures, as in conventional electronics, but also on the spin of those carriers. This provides a further degree of freedom to this nascent form of electronics, exploiting the spin dependence of conduction within magnetic nanostructures. This new electronics came into being only recently, with the discovery of giant magnetoresistance in metallic magnetic multilayers (Fe/Cr) by A. Fert and coworkers (Orsay, France) [95] and P. Grünberg and coworkers (Jülich, Germany) [96] in 1988.

Metallic magnetic multilayers are stacks of alternately ferromagnetic and non-magnetic layers with individual thicknesses of nanometer order. In Fe/Cr multilayers, for certain Cr thicknesses, there is indirect antiferromagnetic exchange coupling in such a way that the magnetisation of two adjacent Fe layers are antiparallel when there is no magnetic field. When a magnetic field is applied, this coupling can be overcome and the magnetisations of all the Fe

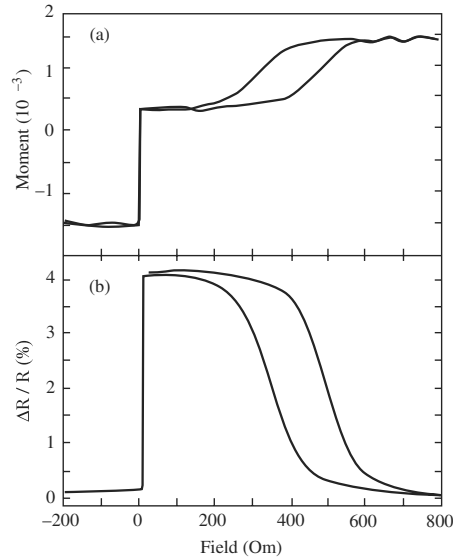
layers are aligned. The resistance of the multilayer then varies significantly as shown in Fig. 14.27. This variation is referred to as giant magnetoresistance (GMR). GMR is defined as the ratio of the resistance variation in a magnetic field to the resistance in the parallel state, i.e.,

$$\text{GMR} = \frac{R_{\text{AP}} - R_{\text{P}}}{R_{\text{P}}},$$

where  $R_{\text{P}}$  and  $R_{\text{AP}}$  are the resistances measured in the parallel (P) and antiparallel (AP) configurations of the magnetisations. This variation reached values of 80% in these first measurements made on multilayers (Fe/Cr), but values of 220% have since been reported in this same system [97]. The property has subsequently been obtained in many systems of the form (F/NM) $_n$ , where F is a ferromagnetic layer of some transition metal, such as Fe, Co, Ni or an alloy of these, and NM is a non-magnetic metal, such as a transition metal (Cr, Ru, etc.) or a rare metal (Cu, Au, Ag). Similar effects have also been observed in multilayers in which the antiparallel configuration of the magnetisations is implemented by a difference in the coercive fields of adjacent ferromagnetic layers due to the use of two different ferromagnetic materials, or the same material with different thicknesses.

But the best known structure is undoubtedly the spin valve structure introduced by B. Dieny and coworkers [98]. It comprises a soft magnetic layer separated by a non-magnetic layer from a hard magnetic layer pinned by exchange coupling with an antiferromagnetic material such as NiO or a ferromagnetic material such as FeMn. The variations of the magnetisation and resistance in a spin valve structure are shown in Fig. 14.28. When the magnetic field goes from negative to positive values, the magnetisation of the free permalloy layer (NeFe) suddenly reverses in a very weak positive field (see Fig. 14.28a), whereas the magnetisation of the pinned layer remains fixed. This results in a sudden change in the resistance of the structure (see Fig. 14.28b). The steep slopes obtained in the variation of the resistance with changing field are currently used in many applications such as magnetic field sensors (originally commercialised by Honeywell in 1994) and read heads for hard disks (first sold by IBM in 1997), but also in non-volatile magnetic memories (magnetic random access memories MRAM) which are expected to replace Si-based memories (Motorola, Infineon, IBM, 2004).

The giant magnetoresistance can be measured by applying a current in the plane of the layers (current-in-plane geometry, CIP), as was done in the first measurements on Fe/Cr multilayers, or perpendicular to the plane of the layers (current perpendicular to the plane, CPP). The first measurements in CPP geometry were carried out at Michigan State University [99] on multilayers sandwiched between two superconducting Nb layers. Other methods, the fabrication of nanowires [100, 101], or oblique deposition on prepatterned substrates [102], were subsequently developed to allow measurements to be made in this geometry. Figure 14.29 shows cross-sections of a nanowire and a

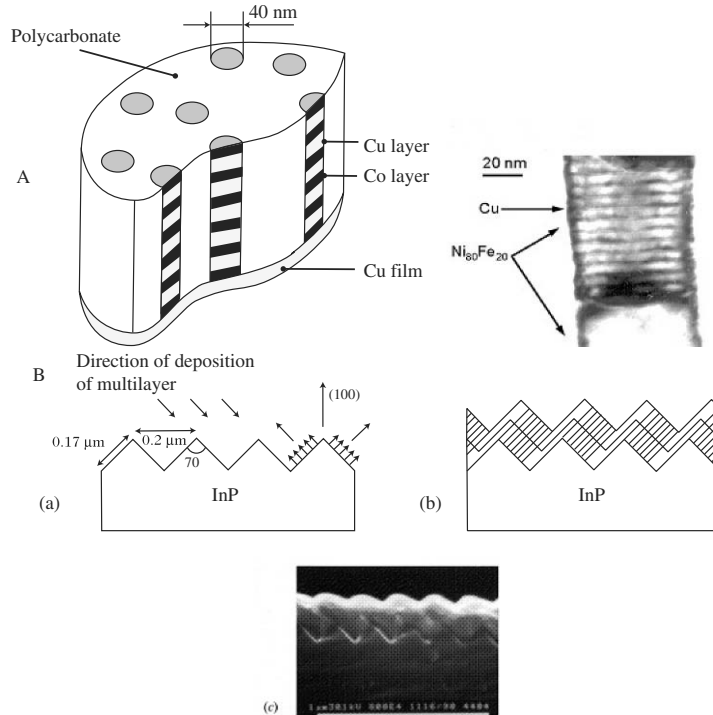


**Fig. 14.28.** (a) Magnetisation and (b) magnetoresistance curves for an NiFe/Cu/NiFe/FeMn spin valve structure [98] in which the magnetisation of one of the NiFe layers is pinned by contact with an FeMn layer. Very steep slopes are thereby obtained in the resistance curves

structure deposited on a prepatterned substrate, as observed by transmission electron microscopy.

Figure 14.30 shows the variation of the GMR of the Co/Cu system as a function of the non-magnetic thickness of Cu in the two geometries. The variation of the GMR with the thickness of the non-magnetic layers is oscillatory [103] due to the oscillatory nature of the indirect exchange coupling between the ferromagnetic layers. It is clear from these variations that the effect of the GMR is greater and more persistent for greater thicknesses in the perpendicular geometry. These differences arise in part from the two different length scales in the two geometries. While the length scale in the CIP geometry is the mean free path  $\lambda$ , i.e., the average distance travelled by electrons between two consecutive collisions, which is of the order of a few nanometers to several tens of nanometers, the length scale in the CPP geometry is the spin diffusion length  $l_{sf}$ , which is the average distance over which the electron conserves its spin, something like ten times longer than  $\lambda$ . This difference of length scale arises from spin accumulation effects present in the CPP geometry (see Sect. 14.2.2).

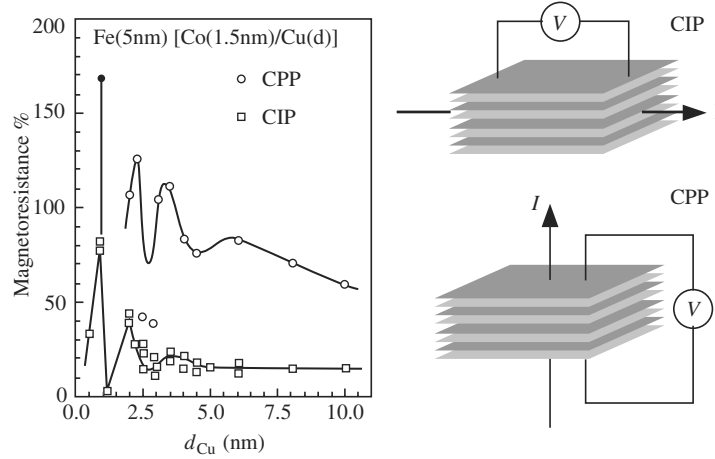
A further consequence of spin accumulation is the possibility of reversing the magnetisation by injecting spin into an  $F_1(t_{F_1})/NM/F_3(t_{F_2})$  structure, in which  $F_1$  and  $F_2$  are thin layers of ferromagnetic metals with thicknesses  $t_{F_1}$  and  $t_{F_2}$ , respectively (where  $t_{F_1} \ll t_{F_2}$ ) and NM is a non-magnetic metal. This



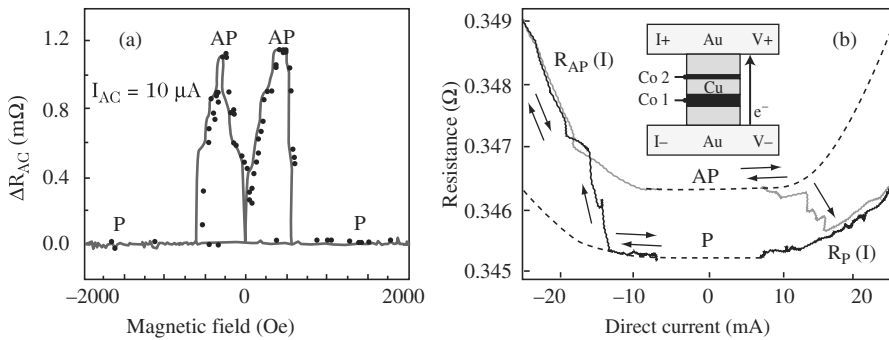
**Fig. 14.29.** Cross-sectional view by transmission electron microscopy of (A) a nano-wire obtained by electrodeposition in an ion-etched polycarbonate membrane and (B) a ‘factory roof structure’ obtained by oblique deposition of the multilayer on a prepatterned substrate. In this case the GMR in CIP and CPP geometries can be measured on the same sample when the current is parallel or perpendicular to the factory roof

phenomenon in which the magnetisation is reversed by a polarised current was predicted by Slonczewski in 1996 [104]. However, it has only recently been observed [105, 106].

Figure 14.31a shows the GMR in CPP geometry for a small alternating current of  $10\ \mu\text{A}$  passing through a  $\text{Co}(15\ \text{nm})/\text{Cu}(10\ \text{nm})/\text{Co}(2.5\ \text{nm})$  pillar. In this structure, the thick Co layer has the task of polarising the spin of the current injected into the structure, and it is the thin layer which reverses under the effect of the polarised current. Figure 14.31b shows the variation in the resistance of this same pillar when a large direct current  $I_{\text{DC}}$  is injected. Starting from  $I_{\text{DC}} = 0$ , in the parallel magnetisation configuration, when a negative current is injected, the system remains in the parallel state (corresponding to a resistance  $R_{\text{P}}$ ), until the applied current reaches a critical value  $I_{\text{AP}} = -I_{\text{c}}$  at which it flips into the antiparallel state of resistance  $R_{\text{AP}}$ . Likewise, if a positive current is now applied, the system remains in its antiparallel state



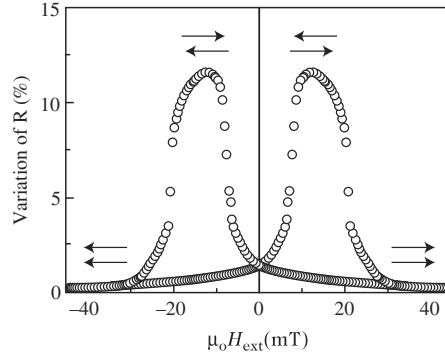
**Fig. 14.30.** Variation of the GMR of Co/Cu multilayers with the non-magnetic thickness of Cu in CIP and CPP geometries [99]



**Fig. 14.31.** (a) Magnetoresistance curve of a  $200 \times 600 \mu m^2$  Co/Cu/Co pillar for a current  $I_{AC} = 10 \mu A$ . Measurements were made at 4 K. (b) Resistance of the pillar as a function of the direct current applied to it [106]. Starting from the state P at  $I = 0$ , the magnetisation of the thin layer remains parallel to that of the thick layer until the current density in the pillar reaches a critical value  $-I_c$ , at which point the thick layer flips over and an antiparallel configuration is obtained. This is revealed by a high resistance. Likewise, if a positive current is now injected into the pillar, the structure remains in an antiparallel state of the magnetisations up to a value  $+I_c$ , at which point the magnetisation of the thin layer flips over and a parallel state is obtained, as revealed by a low resistance

( $R = R_{AP}$ ) until the current reaches the critical value  $I_P = I_c$ , at which it flips into the parallel state characterised by resistance  $R_P$ . This type of effect could have very wide applications in switching very small magnetic devices.

Since the observation of a very high tunnel magnetoresistance at room temperature by Moodera and coworkers [107] in 1995, a great deal of research has



**Fig. 14.32.** Room temperature magnetoresistance curve of a CoFe/Al<sub>2</sub>O<sub>3</sub>/Co tunnel junction [107]

been devoted to this new form of spin-dependent transport. Tunnel magnetoresistance (TMR) is observed in magnetic tunnel junctions, which are made from two conducting ferromagnetic layers separated by a nanoscale insulating layer. Figure 14.32 shows an example of such a high magnetoresistance measured at room temperature in a CoFe/Al<sub>2</sub>O<sub>3</sub>/Co trilayer. As in the GMR effect, TMR results from a variation in the resistance when the magnetisation configurations in the two ferromagnetic layers is changed from antiparallel to parallel by applying a magnetic field, although in this case electron transport occurs by the tunnel effect [108]. In this example, the TMR, defined as the ratio of the variation in the resistance to the resistance in the parallel state, viz.,

$$\text{TMR} = \frac{R_{\text{AP}} - R_{\text{P}}}{R_{\text{P}}},$$

is 12%. The possibility of having a spin-dependent tunnel effect in structures involving two ferromagnetic electrodes was demonstrated by Jullière in 1975. In the Jullière model [109], the TMR is related to the spin polarisations  $SP_1$  and  $SP_2$  by

$$\text{TMR} = \frac{R_{\text{AP}} - R_{\text{P}}}{R_{\text{P}}} = \frac{2SP_1SP_2}{1 - SP_1SP_2},$$

where

$$SP_i = \frac{D_{\uparrow}(\varepsilon_{\text{F}}) - D_{\downarrow}(\varepsilon_{\text{F}})}{D_{\uparrow}(\varepsilon_{\text{F}}) + D_{\downarrow}(\varepsilon_{\text{F}})}$$

is proportional to the difference in the densities of states at the Fermi level,  $D_{\sigma}(\varepsilon_{\text{F}})$ , for majority spin electrons ( $\sigma = \uparrow$ ) and minority spin electrons ( $\sigma = \downarrow$ ). A more detailed description of this model is given in Sect. 14.2.3, but it is clear that 100% polarised materials should lead to very high tunnel magnetoresistances (indeed, theoretically infinite).

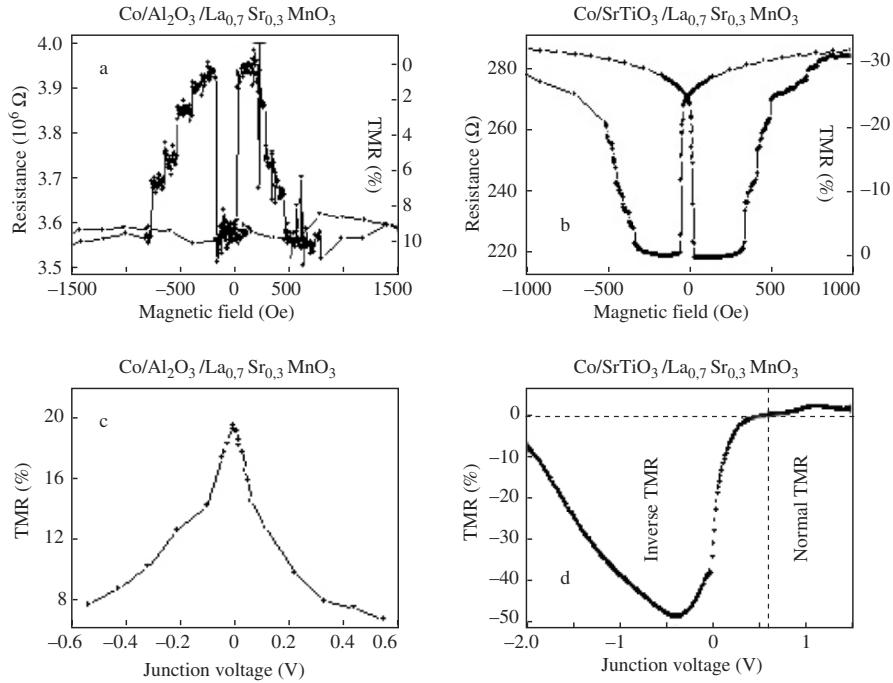
The use of half-metallic materials, i.e., conducting for one spin direction and having a gap at the Fermi level (insulating) for the other spin direction, hence exhibiting an almost total spin polarisation, has established record magnetoresistances of 1 800% [110] with the half-metal  $\text{La}_{0.7}\text{Sr}_{0.3}\text{MnO}_3$  at low temperatures. However, from the point of view of applications, transition metals with spin polarisations of around 70% (corresponding to TMR values of about 50%) are actually used. Half-metals with very high spin polarisations at room temperature have not yet proven their worth.

The role played by the barrier in the tunnel effect has been demonstrated by comparing results obtained with  $\text{LaSrMnO}_3/\text{SrTiO}_3/\text{Co}$  and  $\text{LaSrMnO}_3/\text{Al}_2\text{O}_3/\text{Co}$  junctions, which are illustrated in Fig. 14.33 [111]. The positive TMR obtained with an alumina barrier ( $\text{Al}_2\text{O}_3$ ) reflects the positive polarisation at the  $\text{Co}/\text{Al}_2\text{O}_3$  interface. With a  $\text{Co}/\text{SrTiO}_3$  interface, the TMR is inverted (lower resistance in the antiparallel state), indicating a negative polarisation at the interface. Different barriers also lead to very different dependences of the TMR on the voltage applied to the junction, as can be seen from the results obtained on these two types of junction (Figs. 14.33c and d). These differences in behaviour have been attributed to the fact that it is *s* electrons with positive polarisation that are responsible for the tunnel effect when the barrier is made of  $\text{Al}_2\text{O}_3$ , whereas it is *d* electrons with negative polarisation that are transmitted by the tunnel effect when the barrier is made of  $\text{SrTiO}_3$ . These differences in the polarisation and the voltage dependence have been linked to the bonds forming at the electrode/insulator interface [112].

Today's spin electronics, based purely on metallic materials, has provided a new way of storing and reading magnetic data. Spin electronics based on semiconductors could combine storage, detection, and logic elements to provide new multifunctional components. Semiconductors can also exploit the advantages of a much longer spin lifetime [113] than in conducting materials, and novel effects due to quantised levels existing in quantum wells, or the possibility of transforming magnetic data into an optical signal. This explains the many attempts by research teams to combine ferromagnetic and semiconducting materials.

However, progress in this area has been slow due to the problem of injecting spin into a metal/semiconductor interface. The results obtained with structures in which spin is injected into a semiconductor from a ferromagnetic metal have demonstrated that injection at the ferromagnetic metal/semiconductor interface is totally inefficient [114, 115]. This difficulty has recently been moderated by using diluted magnetic semiconductors [116, 117]. In this case, an injection efficiency of up to 90% has been measured by determining the light polarisation. At the present time, very few experiments have been carried out using a ferromagnetic semiconductor as spin polariser and analyser [118], but this line of investigation will certainly be exploited in the future and will lead to new components for electronics, such as the spinFET (spin field effect transistor) proposed by Datta and Das (see Fig. 14.34) [119]. As far as





**Fig. 14.33.** Magnetoresistance curves at 4 K for (a)  $\text{La}_{0.7}\text{Sr}_{0.3}\text{MnO}_3/\text{Al}_2\text{O}_3/\text{Co}$  and (b)  $\text{La}_{0.7}\text{Sr}_{0.3}\text{MnO}_3/\text{SrTiO}_3/\text{Co}$  tunnel junctions. The positive polarisation at the  $\text{Co}/\text{Al}_2\text{O}_3$  interface gives rise to a normal TMR with higher resistance in the antiparallel state. In contrast, the negative polarisation of the  $\text{Co}/\text{SrTiO}_3$  interface generates an inverse TMR with lower resistance in the antiparallel state. Variation of the TMR with applied bias at the junction for (c)  $\text{La}_{0.7}\text{Sr}_{0.3}\text{MnO}_3/\text{Al}_2\text{O}_3/\text{Co}$  and (d)  $\text{La}_{0.7}\text{Sr}_{0.3}\text{MnO}_3/\text{SrTiO}_3/\text{Co}$  junctions. The voltage behaviour of the tunnel junctions is very different depending on the type of barrier used. Such differences can be interpreted in terms of the densities of states of the electrodes weighted by the effect of the tunnel barrier

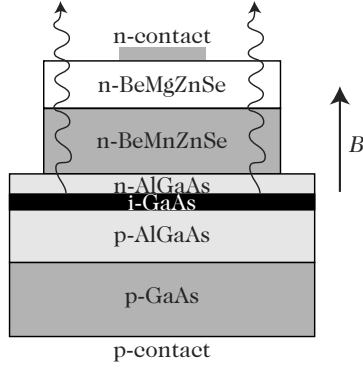
applications are concerned, it remains to find ferromagnetic semiconductors at room temperature [120].

## 14.2.2 Origins and Mechanisms of Spin Electronics

### Transport in Metals

#### *Transport in Ferromagnetic Metals. Two-Current Model*

As discussed in Sect. 14.2.1, the ferromagnetic transition metals Fe, Co, and Ni and their alloys which are generally used as magnetic layers in spin electronics fulfil the Stoner criterion. The  $3d_{\uparrow}$  and  $3d_{\downarrow}$  subbands are therefore shifted by the presence of an exchange interaction within these materials. This shift



**Fig. 14.34.** New components can be designed by integrating metals and semiconductors into spin electronics. The figure shows a spinFET proposed by Datta and Das [119]

not only generates a spontaneous magnetisation, but also different densities of states at the Fermi level and different mobilities for electrons with the two spin directions. This results in a current with spin polarisation which can be used in the tunnel effect of magnetic tunnel junctions (TMR) and an asymmetry of the resistivity with respect to the spin which is exploited in magnetic multilayers and gives rise to the GMR effect.

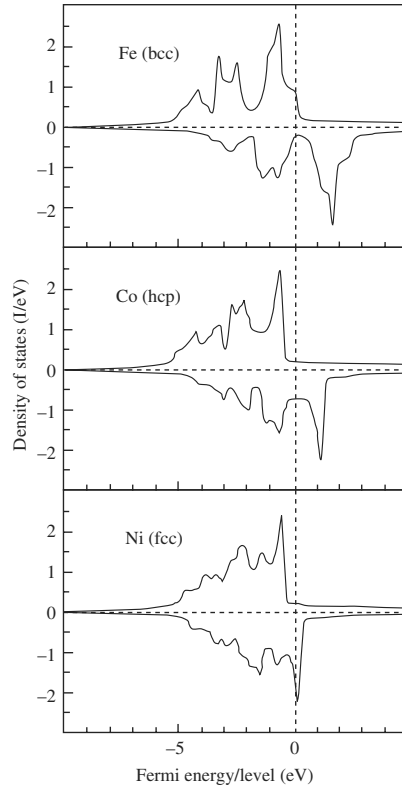
#### *Two-Current Model and Mott Model*

The Mott model [27] assumes that, in the transition metals, the  $d$  electrons are responsible for magnetism, the  $s$  electrons are responsible for conduction, and resistivity is due to  $s \rightarrow d$  transitions. The lack of symmetry between the densities of states at the Fermi level for spin  $\uparrow$  and spin  $\downarrow$  states will thus lead to different transition probabilities for spin  $\uparrow$  and spin  $\downarrow$   $s$  electrons. In the case of Co and Ni (see Fig. 14.35), the filled  $3d_{\uparrow}$  subband is located below the Fermi level and the density of states at the Fermi level in the majority band, viz.,  $D_{\uparrow}(\varepsilon_F)$ , is thus extremely small, whereas the minority subband crosses the Fermi level in such a way that its density of states  $D_{\downarrow}(\varepsilon_F)$  is large. The  $s \rightarrow d$  transition probabilities are therefore very different for spin  $\uparrow$  electrons and spin  $\downarrow$  electrons.

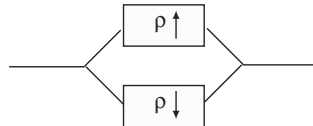
In the Mott model, at low temperatures, when electron–magnon collisions with spin reversal are frozen, the spin  $\uparrow$  and spin  $\downarrow$  conduction electrons (see Fig. 14.36) will lead the current in parallel into two independent channels with different resistivities

$$\rho_{\sigma} = \frac{m_{\sigma}}{n_{\sigma} e^2 \tau_{\sigma}} = \frac{1}{n_{\sigma} e \mu_{\sigma}},$$

where  $m_{\sigma}$ ,  $n_{\sigma}$ ,  $\tau_{\sigma}$  and  $\mu_{\sigma}$  are the effective mass, the electron number per unit volume, the relaxation time and the mobility of electrons with spin  $\sigma$ , respectively. The relaxation time is related to the scattering potential  $V_{\sigma}$  and the density of states at the Fermi level  $D_{\sigma}(E_F)$  by



**Fig. 14.35.** Density of states of Fe, Co and Ni



**Fig. 14.36.** According to the Mott model, at low temperatures, conduction inside a ferromagnetic transition metal is due to two independent channels of spin  $\uparrow$  and spin  $\downarrow$  electrons. This can be represented by a circuit diagram with two resistances  $\rho_{\uparrow}$  and  $\rho_{\downarrow}$  in parallel. The resistivity of the material will thus be  $1/\rho = 1/\rho_{\uparrow} + 1/\rho_{\downarrow}$

$$\frac{1}{\tau_{\sigma}} \approx |V_{\sigma}|^2 D_{\sigma}(\varepsilon_F) .$$

The resistivity  $\rho$  of the ferromagnetic metal (see Fig. 14.36) is therefore

$$\rho = \frac{\rho_{\uparrow}\rho_{\downarrow}}{\rho_{\uparrow} + \rho_{\downarrow}} .$$

The spin asymmetry of the conduction in the two channels is characterised by the spin asymmetry coefficients

$$\alpha = \frac{\rho_{\downarrow}}{\rho_{\uparrow}} \quad \text{or} \quad \beta = \frac{\rho_{\downarrow} - \rho_{\uparrow}}{\rho_{\downarrow} + \rho_{\uparrow}} .$$

The further  $\alpha$  ( $\beta$ ) is from 1 (0), the more pronounced the spin asymmetry becomes. The spin dependence of  $\rho_{\sigma}$  thus has an intrinsic origin connected to the spin dependence of  $m_{\sigma}$ ,  $n_{\sigma}$ , and  $D_{\sigma}(\varepsilon_{\text{F}})$  for the ferromagnetic metal. For example, due to the large difference in the density of states at the Fermi level, with  $D_{\downarrow}(\varepsilon_{\text{F}}) \gg D_{\uparrow}(\varepsilon_{\text{F}})$  in the case of Ni or Co shown in Fig. 14.35, systems based on Ni or Co will tend to have  $\rho_{\downarrow} > \rho_{\uparrow}$  [122,123] and hence  $\alpha > 1$ .

There is also an extrinsic origin for the spin dependence of  $\rho_{\sigma}$  which is connected with the spin dependence of scattering by defects or impurities and which is reflected in the spin dependence of the scattering potential  $V_{\sigma}$ . For example, with 1% of Fe impurities in Ni, the ratio  $\alpha$  is 20, whereas it is 0.45 for 1% of Cr impurities in Ni [122]. These values, greater than or less than unity, are explained by the modification of the density of states at the site of the impurity. By inserting Co impurities in Ni, which is a nearby element in the classification of the periodic table, the density of states at the site of the impurity is only slightly modified and  $\alpha$  remains greater than unity. In the case of a Cr impurity, given that Cr is more distant from the element (Ni) in the matrix according to the periodic classification, the  $d$  states of the impurity can no longer hybridise with the  $d$  states of the matrix. A virtual bound state then forms just above the Fermi level in the majority band by hybridisation with the  $s$  states, and  $\alpha$  is less than unity. The values of many spin asymmetry coefficients for dilute impurities can be found in [122].

At higher temperatures, electron–magnon collisions cause the two currents to mix. The resistivity then takes the form

$$\rho = \frac{\rho_{\uparrow}\rho_{\downarrow} + \rho_{\uparrow\downarrow}(\rho_{\uparrow} + \rho_{\downarrow})}{\rho_{\uparrow} + \rho_{\downarrow} + 4\rho_{\uparrow\downarrow}} ,$$

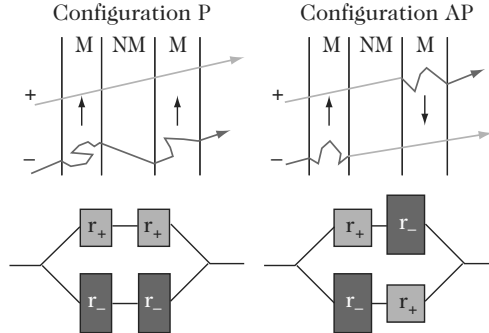
where  $\rho_{\uparrow\downarrow}$  is a term accounting for the mixing of the two currents [124].

The mechanism for giant magnetoresistance is the extrapolation of the two-current model to the case of multilayers.

## Giant Magnetoresistance

### *GMR Mechanism*

Figure 14.37 illustrates the mechanism for GMR at low temperatures where spin-reversing scattering processes are unlikely, in the case where  $\alpha$  is much greater than unity, i.e., electrons with majority spin are not much scattered. In the parallel configuration (P), when the magnetisations of all layers are aligned, electrons in the + spin channel (– spin channel) are majority (minority) spin electrons in all the layers, hence weakly (strongly) scattered throughout the structure. This situation is characterised by different resistances  $r_{+}$  and  $r_{-}$  for the two channels and a resistance



**Fig. 14.37.** Mechanism for conduction in a multilayer for the two spin directions  $S_z = +1/2$  (+ channel) and  $S_z = -1/2$  (– channel). *Top left:* Configuration of parallel magnetisations (P). *Top right:* Configuration of antiparallel magnetisations (AP). *Bottom:* Equivalent circuit diagrams. Scattering is indicated by a break in the trajectory of conduction electrons with majority spin (spin  $\uparrow$ ) and minority spin (spin  $\downarrow$ )

$$r_P = \frac{r_+ r_-}{r_+ + r_-}$$

in the parallel state. When the spin asymmetry is very pronounced, i.e.,  $r_+ \ll r_-$ , fast electrons cause a short-circuit effect (see the equivalent circuit diagram, bottom left in Fig. 14.37) and the resistance measured in the parallel state is low and equal to  $r_+$ . In the antiparallel configuration of the magnetisations, since each spin direction is alternately the majority then the minority spin, the electrons with each spin direction are alternately weakly then strongly scattered. This leads to an averaging effect in each channel and the resulting resistance

$$r_{AP} = \frac{r_+ + r_-}{4}$$

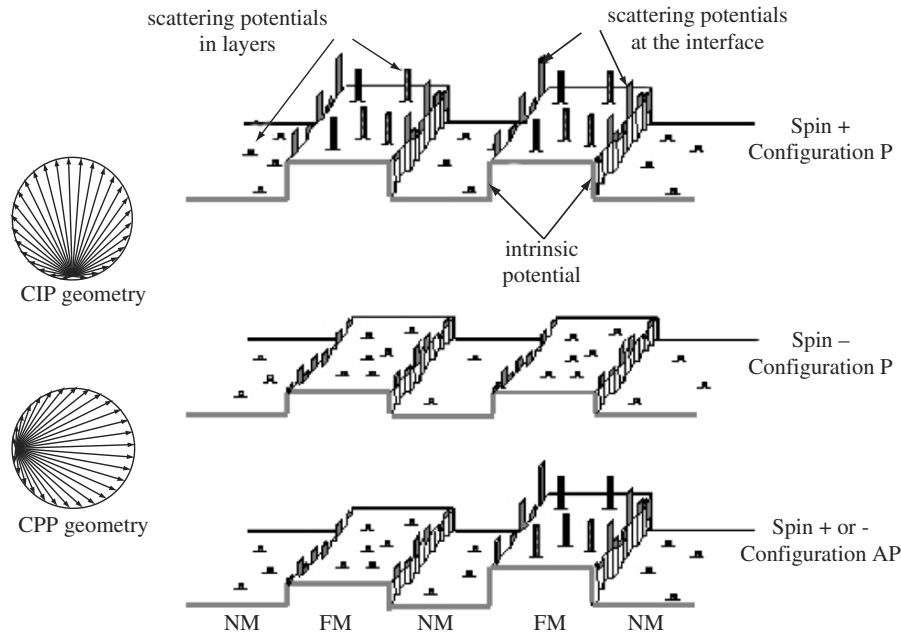
is higher than in the parallel configuration. The magnetoresistance follows from these expressions:

$$\text{GMR} = \frac{r_{AP} - r_P}{r_P} = \frac{(r_- - r_+)^2}{4r_+ r_-}.$$

This image is valid in both CIP and CPP geometries provided that the thicknesses are small compared with the appropriate length scale, i.e., compared with the mean free path  $\lambda$  in CIP geometry and the spin diffusion length  $l_{sf}$  in CPP geometry.

#### *Microscopic Origins of GMR*

The microscopic origins of GMR can be traced back to the potential landscape seen by the conduction electrons, as depicted in Fig. 14.38. This landscape contains two types of potential:



**Fig. 14.38.** *Right:* Potential landscapes seen by conduction electrons within a multilayer. *Left:* Schematic current distribution in  $k$  space for CIP and CPP geometries. The electrons will manifest different sensitivities to the various types of potential in these two geometries

- a periodic intrinsic potential representing the perfect multilayer,
- extrinsic potentials associated with the presence of impurities in the layers and the roughness of the interfaces.

Spin + and spin – refer to the spin orientations in an absolute reference frame. In the configuration where the magnetisations of all layers are parallel, spin + (–) is the direction of the majority (minority) spin in all the layers. In the antiparallel configuration, + and – are alternately the majority and minority spin directions owing to the alternating magnetisation directions.

In Fig. 14.38, the intrinsic potential is shown by steps. The difference in height of the steps for majority and minority spin electrons arises from the shift due to exchange between spin  $\uparrow$  and spin  $\downarrow$  spin states in a ferromagnetic metal. This potential is periodic for a periodic multilayer. In the parallel configuration (P) of the magnetisations, the heights of the potential are different for electrons of spins + and – in a given spin channel, but they are the same in all the layers. In the antiparallel configuration (AP), since each spin direction is alternately majority then minority, small and large steps alternate for each of the spin directions. This intrinsic potential generates wavefunctions with different superlattices for the two channels in the parallel and

antiparallel configurations in Fig. 14.38, and thus leads to a GMR effect even in the absence of spin-dependent scattering.

The second contribution, due to extrinsic potentials, arises from the presence of defects (impurities in the layers or interface roughness). They are represented by narrow peaks in Fig. 14.38. In ferromagnetic materials, as the scattering is spin-dependent, the scattering potentials have different heights for spin  $\uparrow$  and spin  $\downarrow$ .

The precise manner in which the electrons will react to these two types of potential will depend on the geometry under consideration [125]. In CPP geometry, electrons travel through all the layers and see all the types of potential. In this geometry, the intrinsic potentials will intervene through a spin-dependent interface resistance. In CIP geometry, the role of the intrinsic potentials is to channel the electrons in certain layers so that the electrons will probe the potentials at the interfaces and in the bulk of the magnetic and non-magnetic layers in different ways.

#### *Camley–Barnas Model for GMR in CIP Geometry*

The models due to Camley and Barnas [126], Johnson and Camley [127], and Barnas et al. [128] are based on the phenomenological theory of Fuchs and Sondheimer [129], taking into account the spin-dependent scattering at the interfaces [126], in the bulk of the layers [127], or both types of scattering [128], respectively.

In this approach, electrons with spin  $\sigma$  are described by their distribution function  $f^\sigma(\mathbf{k}, \mathbf{r})$ , which represents the probability that an energy state  $E$  is occupied. For thin films and multilayers, the distribution function depends on the position  $r$  of the electron [129]. At equilibrium at temperature  $T$ , i.e., in the absence of any electric field, the distribution function is determined by Fermi–Dirac statistics:

$$f_0(\mathbf{k}) = \frac{1}{\exp \frac{E - \varepsilon_F}{k_B T} + 1} .$$

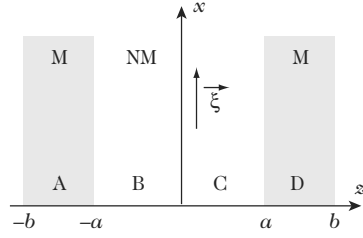
Under the action of an electric field  $\boldsymbol{\xi}$ , the distribution function (like the Fermi sphere) moves away from its equilibrium value. It can then be expressed in the form

$$f^\sigma(\mathbf{k}, \mathbf{r}) = f_0(\mathbf{k}) + g^\sigma(\mathbf{k}, \mathbf{r}) ,$$

where  $g^\sigma(\mathbf{k}, \mathbf{r})$  is the deviation from the equilibrium distribution.

In the following, we consider an electric field applied in the  $x$  direction, with the  $z$  axis normal to the plane of the layers, as shown in Fig. 14.39. By differentiation, we obtain

$$df^\sigma = \frac{\partial f^\sigma}{\partial t} dt + \frac{\partial f^\sigma}{\partial k_x} dk_x + \frac{\partial f^\sigma}{\partial z} dz ,$$



**Fig. 14.39.** Notation for the structure under consideration, formed from two magnetic layers M separated by a non-magnetic layer NM. The electric field is applied in the  $x$  direction and the  $z$  axis is perpendicular to the plane of the layers

whence

$$\frac{df^\sigma}{dt} = \frac{\partial f^\sigma}{\partial t} + \frac{\partial f^\sigma}{\partial k_x} \frac{dk_x}{dt} + \frac{\partial f^\sigma}{\partial z} \frac{dz}{dt}.$$

The term  $dk_x/dt$  is easily deduced from Newton's second law:

$$\frac{dk_x}{dt} = -\frac{e\xi}{\hbar}.$$

In the free electron approximation, viz.,

$$E = \frac{\hbar^2 k^2}{2m},$$

and neglecting second order terms, the equation becomes

$$\frac{df^\sigma}{dt} = \frac{\partial f^\sigma}{\partial t} - \frac{\partial f_0^\sigma}{\partial E} ev_x \xi + \frac{\partial g}{\partial z} v_z.$$

In the relaxation time approximation,  $f$  relaxes exponentially to its equilibrium value  $f_0$  ( $g$  relaxes exponentially to zero) in a characteristic time  $\tau$  (relaxation time):

$$\frac{df^\sigma}{dt} = \frac{dg^\sigma}{dt} = -\frac{g^\sigma}{\tau}.$$

Substituting this in, we obtain

$$-\frac{g^\sigma}{\tau} = \frac{\partial f^\sigma}{\partial t} - \frac{\partial f_0^\sigma}{\partial E} ev_x \xi + \frac{\partial g}{\partial z} v_z.$$

In the stationary regime, the distribution function does not vary in time. The various terms contributing to its modification must therefore balance one another:

$$-\frac{g^\sigma}{\tau} = -\frac{\partial f_0^\sigma}{\partial E} ev_x \xi + \frac{\partial g}{\partial z} v_z.$$

In this equation, the variation of the distribution function due to the application of an electric field (first term on the right) and the variation due to spatial



inhomogeneity in the  $z$  direction are balanced by the scattering phenomena which ensure return to equilibrium.

Solutions take the form

$$g^\sigma(z, k) = v_x e E \frac{df_0}{dE} \left( 1 + A^\sigma e^{-z/\tau v_z} \right) .$$

The constants  $A$  in the different regions are related by the boundary conditions. For the surfaces, if  $R$  is the probability of specular reflection:

$$\begin{aligned} g_{A+}^\sigma &= R g_{A-}^\sigma & \text{at } z = -b , \\ g_{D-}^\sigma &= R g_{D+}^\sigma & \text{at } z = b . \end{aligned}$$

Indices  $+$  and  $-$  refer to the positive ( $+$ ) and negative ( $-$ ) components of the velocity in the  $z$  direction. At the interfaces, the electron of spin  $\sigma$  has probability  $R_\sigma$  of being reflected and probability  $T_\sigma$  of being transmitted. The boundary conditions are thus, for example, for the interface between regions  $A$  and  $B$  at  $z = -a$ ,

$$\begin{aligned} g_{A-}^\sigma &= T_\sigma g_{B-}^\sigma + R_\sigma g_{A+}^\sigma , \\ g_{B+}^\sigma &= T_\sigma g_{A+}^\sigma + R_\sigma g_{B-}^\sigma . \end{aligned}$$

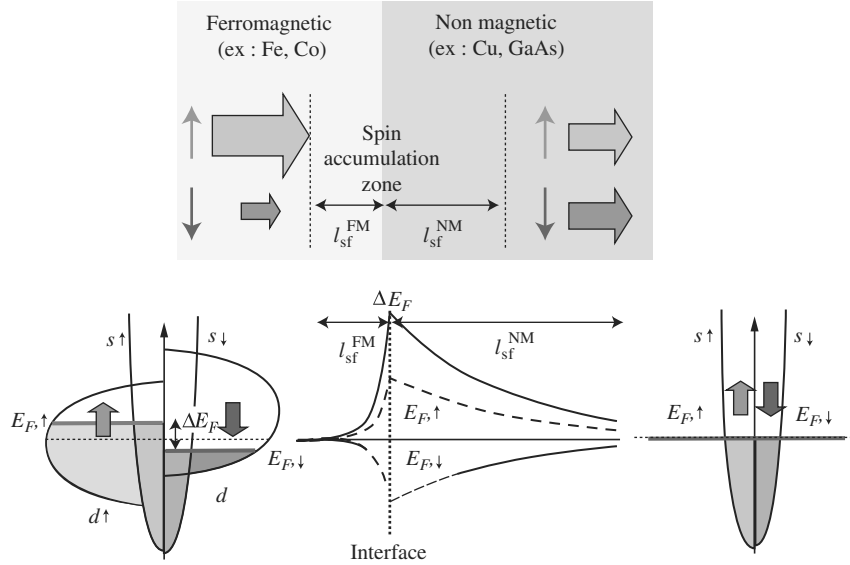
From the expressions for  $g$  in the various spatial regions, one may calculate the current density:

$$j_x(z) = -2e \left( \frac{m}{h} \right)^3 \int v_x g(v, z) d^3v .$$

Using such models, one can describe the variation of the effect with number of layers and the mean free path over the thickness of the layers, and one can deduce the spin asymmetry of scattering.

#### *Spin Accumulation and GMR in CPP Geometry*

Consider the situation illustrated in Fig. 14.40, which shows an interface between a ferromagnetic material and a non-magnetic material. In the ferromagnetic material, the  $\uparrow$  spins represent the majority, i.e., there are more spin  $\uparrow$  than spin  $\downarrow$ . In the non-magnetic material, the two spin populations are identical. When they transit from the ferromagnetic layer to the non-magnetic layer, the  $\uparrow$  spins will thus accumulate at the FM/NM interface, before inverting in the spin  $\downarrow$  channel. This accumulation occurs over a length which is the spin diffusion length  $l_{sf}$ . The result is a shift in the Fermi levels of the two spin directions, denoted  $\Delta E_F$ , which extends over a distance  $l_{sf}$  as shown in Fig. 14.40b. This spin diffusion length is the average distance travelled by a conduction electron between collisions reversing its spin. It can be expressed in terms of the mean free path  $\lambda$  and the spin mean free path



**Fig. 14.40.** Illustration of the spin accumulation effect at a ferromagnetic (FM)/non-magnetic (NM) interface. The majority spins accumulate before flipping over in the minority spin channel. This accumulation causes a shift  $\Delta E_F$  in the Fermi levels of the two spin populations, which extends over a distance  $l_{sf}$  on either side of the interface

$$\lambda_s = \sum_{\text{between two spin reversals}} \lambda,$$

which is the total mean distance travelled by the electron between two spin reversals:

$$l_{sf} = \sqrt{\frac{\lambda_s \lambda}{3}}.$$

These quantities are shown in Fig. 14.41.

This spin accumulation at the FM/NM interface results in a spin polarisation of the current which can be used to reverse the magnetisation of another ferromagnetic layer [105,106], or it can be exploited to inject a polarised current into a semiconductor when impedance-matching conditions are fulfilled at the ferromagnetic metal/semiconductor interface.

### 14.2.3 Magnetoresistance of Tunnel Junctions

In this section, we shall be concerned with the spin-dependent tunnel effect, following the approach of Jullière [109], which allows one to relate the tunnel magnetoresistance to the spin polarisations of the two electrodes. The tunnel

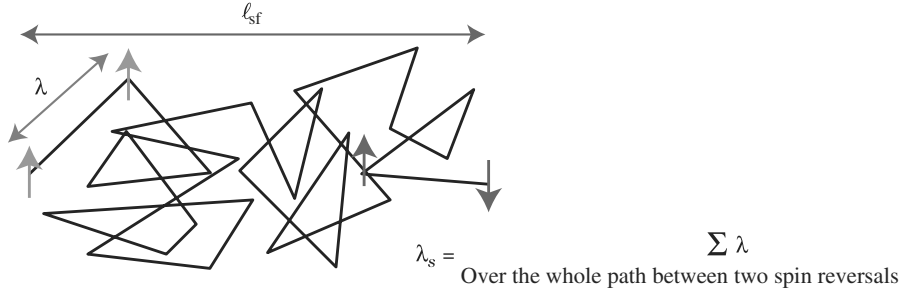


Fig. 14.41. Representation of lengths  $\lambda$  and  $l_{sf}$

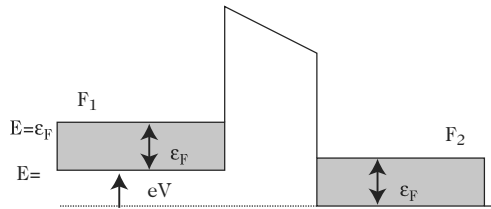


Fig. 14.42. Energy diagram of a metal–insulator–metal junction under a bias  $V$

current takes its origins in the transmission of electrons through a potential barrier when these electrons have energies lower than the barrier height. The energy diagram for such a configuration is shown in Fig. 14.42 for the one-dimensional problem. Electrode  $F_2$  has positive potential relative to electrode  $F_1$ . Shaded regions correspond to levels that are occupied at zero temperature. Applying a voltage across the junction shifts the Fermi levels of the two electrodes, establishing a correspondence between the occupied states of the left-hand electrode  $F_1$  and the empty states of the right-hand electrode  $F_2$ , into which the electrons can then be transmitted.

According to Bardeen [130], the probability of an electron being transmitted from an occupied state  $\psi_1$  of energy  $E$  of the electrode  $F_1$  into an empty state  $\psi_2$  of the same energy in the electrode  $F_2$  can be written in the form

$$P_{12} = \frac{2\pi}{\hbar} |M_{12}|^2 f_1 D_2(E)(1 - f_2).$$

It is the product of two factors: the first is the probability  $f_1$  of the state  $\psi_1$  of the electrode  $F_1$  being occupied by the number  $D_2(1 - f_2)$  of empty states of the electrode  $F_2$ ; the second is the square  $|M_{12}|^2$  of the transfer matrix element, which represents the probability of transmission by the barrier [131]. The current of electrons transmitted by the tunnel effect from electrode  $F_1$  to electrode  $F_2$  at energy  $E$  is obtained by multiplying  $P_{12}$  by the density of states  $\psi_1$  of energy  $E$  in electrode  $F_1$ , which yields

$$I_{1 \rightarrow 2} \propto D_1(E - eV) f_1 |M_{12}|^2 D_2(E)(1 - f_2).$$

where  $D_1$  and  $D_2$  are the densities of states of electrodes  $F_1$  and  $F_2$ , respectively, and  $f$  is the distribution function, i.e., the probability of occupation of a state of energy  $E$  at temperature  $T$ , which is defined by the Fermi–Dirac function.

Using the notation in Fig. 14.42,

$$f_1 = f(E) = \frac{1}{1 + \exp \frac{E - E_F}{k_B T}}$$

and

$$f_2 = f(E + eV) = \frac{1}{1 + \exp \frac{E + eV - E_F}{k_B T}} .$$

In the same way, the current of electrons transmitted by the tunnel effect from electrode  $F_2$  to electrode  $F_1$  is defined by

$$I_{2 \rightarrow 1} \propto D_1(E) [1 - f(E)] |M_{12}|^2 D_2(E + eV) f(E + eV) .$$

The total current is obtained from  $I_{1 \rightarrow 2} - I_{2 \rightarrow 1}$ , integrating over the whole energy range:

$$I(V) \propto \int_{-\infty}^{+\infty} D_1(E) |M_{12}|^2 D_2(E + eV) [f(E) - f(E + eV)] dE .$$

The tunnel current thus depends on the band structure of the electrodes through the densities of states  $D_1$  and  $D_2$ , and the characteristics of the tunnel barrier through  $|M_{12}|^2$ .

Given the factor  $f(E) - f(E + eV)$  which appears in the integrand and which is represented in Fig. 14.43, only those energy states between  $\varepsilon_F - eV$  and  $\varepsilon_F$  can contribute to the tunnel current. This is to be expected, since it is only in this energy range that the filled states of one of the electrodes are found to correspond to the empty states of the other electrode.

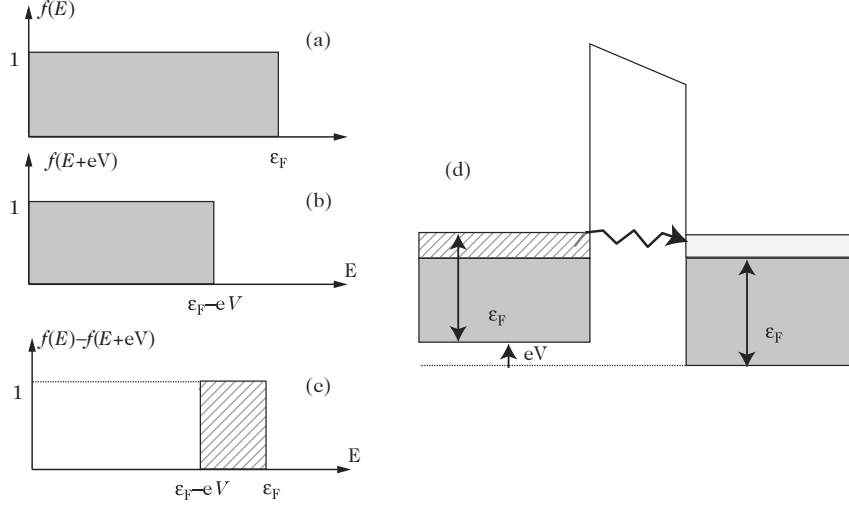
For a low applied bias and treating the transfer matrix element as constant over this energy range, we obtain

$$I \propto D_1(\varepsilon_F) D_2(\varepsilon_F) (eV) .$$

The conductance  $G(V) = I/V$  is thus proportional to the product of the densities of states of the two electrodes, viz.,

$$G(V) \propto D_1(\varepsilon_F) D_2(\varepsilon_F) .$$

To understand the tunnel current between two ferromagnetic electrodes, Jullière [109] takes into account the conservation of electron spin in the tunnel



**Fig. 14.43.** (a) Schematic representation of the distribution functions of the two electrodes for a bias  $V$  across the junction at  $T = 0$ . (c) Schematic representation at  $T = 0$  of the difference  $f(E) - f(E + eV)$  arising in the expression for the tunnel current. (d) Energy diagram for the tunnel junction. Only those states in the energy range  $[\varepsilon_F - eV, \varepsilon_F]$  can contribute to the tunnel effect

process. A spin  $\uparrow$  ( $\downarrow$ ) of the electrode  $F_1$  will thus be transmitted to an empty state of spin  $\uparrow$  ( $\downarrow$ ) of the electrode  $F_2$ .

In the parallel magnetisation configuration, an electron with majority spin  $\uparrow$  in the electrode  $F_1$  will be transmitted to a state with majority spin  $\uparrow$  in the electrode  $F_2$ , as shown in Fig. 14.44. Likewise, an electron with minority spin  $\downarrow$  in the electrode  $F_1$  will be transmitted to a state with minority spin  $\downarrow$  in the electrode  $F_2$ , so that the tunnel current can be written

$$I_P \propto D_{1\text{maj}}D_{2\text{maj}} + D_{1\text{min}}D_{2\text{min}} .$$

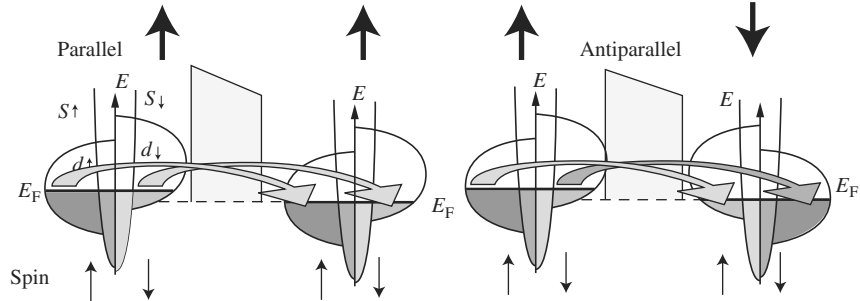
In contrast, in the antiparallel configuration, an electron with majority spin  $\uparrow$  in the electrode  $F_1$  will be transmitted to a state of minority spin  $\downarrow$  in the electrode  $F_2$ . An electron of minority spin  $\downarrow$  in the electrode  $F_1$  will be transmitted to a state of majority spin  $\uparrow$  in the electrode  $F_2$ . The current in the antiparallel configuration will then be

$$I_{AP} \propto D_{1\text{maj}}D_{2\text{min}} + D_{1\text{min}}D_{2\text{maj}} .$$

Finally, the tunnel magnetoresistance is given by

$$\text{TMR} = \frac{R_{AP} - R_P}{R_P} = \frac{I_P - I_{AP}}{I_{AP}} = \frac{2SP_1SP_2}{1 - SP_1SP_2} ,$$

where



**Fig. 14.44.** Schematic representation of the tunnel currents in the parallel configuration (*left*) and antiparallel configuration (*right*) of the magnetisations

**Table 14.3.** Spin polarisation values

		SP
Al <sub>2</sub> O <sub>3</sub> barrier	Ni	+33% [132]
	Ni <sub>80</sub> Fe <sub>20</sub>	+48% [132]
	Co	+42% [132]
	Fe	+45% [132]
	Co <sub>50</sub> Fe <sub>50</sub>	+51% [132]
SrTiO <sub>3</sub> barrier	Co	-24% [111]
	La <sub>0.7</sub> Sr <sub>0.3</sub> MnO <sub>3</sub>	+95% [110]

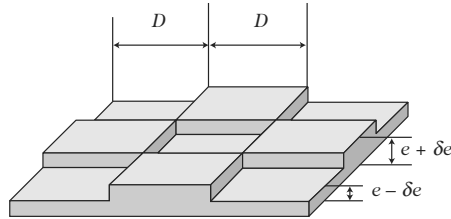
$$SP_i = \frac{D_{i\uparrow}(\varepsilon_F) - D_{i\downarrow}(\varepsilon_F)}{D_{i\uparrow}(\varepsilon_F) + D_{i\downarrow}(\varepsilon_F)}$$

is the spin polarisation of electrode FM<sub>*i*</sub>. This model is commonly used to deduce the spin polarisations from the TMR values. In fact, this polarisation is not an intrinsic property of the material making up the electrode, but depends on the tunnel barrier, or more precisely, the interface [111], as was shown in Sect. 14.2.1. Several values of these polarisations are collected together in Table 14.3.

## Appendix

### A. Nanomagnetism in Real Films: Effect of Defects

No real ultrathin film is ever perfect. It will exhibit crystal defects such as distortion or dislocations, and a certain roughness. Its magnetic properties must therefore be expected to fluctuate in space. Most measurement techniques integrate over areas much greater than the characteristic spatial extent of these fluctuations and thus measure an average value, whose meaning will depend



**Fig. 14.45.** Representation of a rough magnetic film

on the film parameters and the quantity being measured. The discussion here aims to illustrate the general rules governing this effect through a series of examples.

### Roughness and Fundamental Length Scales in Magnetism

A rough magnetic film is often drawn as a regular checkerboard of square terraces, each of side  $D$  and thicknesses alternately  $(e - \delta e)$  and  $(e + \delta e)$ , where  $e$  is the average thickness (see Fig. 14.45). Suppose for example that  $e$  is equal to  $e_c$ , the critical thickness at which a perpendicular easy magnetisation axis flips over to an easy plane (see Sect. 14.1.4). At very large values of  $D$ , at the center of each terrace, the magnetisation is thus either perpendicular or parallel to the film, depending on whether the terrace is less thick or thicker than the critical value, respectively. At the edge of each terrace, the magnetisation rotates in a pseudo-domain wall with lateral extent  $\Delta$ . As in a standard domain wall (see Sect. 14.1.3),  $\Delta$  is determined by competition between the exchange interaction and the magnetocrystalline and dipole anisotropy energies, but it must be calculated specifically in each case.

Suppose now that we measure the way the magnetisation of the film evolves in an external field with equipment which integrates over distances much greater than  $D$ :

- When  $D \gg \Delta$ , each terrace behaves more or less autonomously and the measurement gives an average value of the magnetisation.
- In the other extreme case,  $D \ll \Delta$ , the magnetisations of each terrace are held strictly parallel by the exchange interaction and the measurement characterises a uniform behaviour related to an average value of the magnetic anisotropy.
- The intermediate case,  $D \sim \Delta$  requires a much more involved calculation which can be found in [82], for example. The direction of magnetisation fluctuates from one terrace to another around an average orientation which may be oblique, and the overall measurement of the magnetic anisotropy involves terms of higher order induced by roughness.

A similar argument also applies when the above figure represents a rough intercalating layer between two ferromagnetic layers  $F_1$  and  $F_2$ , coupled by an oscillating interaction through M (see Sect. 14.1.4).

An interesting case discovered by J. Slonczewski in 1991 [83] is one in which the interlayer interaction favours a magnetic configuration in one type of terrace and an antiferromagnetic configuration in the other. Moreover, the thickness  $\Delta$  of the pseudowall must quickly take into account the coupling strength, but the rest of the behaviour is very similar to the magnetic anisotropy case. In the extreme situations  $D \gg \Delta$  and  $D \ll \Delta$ , the measurement averages over the magnetisation and the interlayer interaction, respectively, which tends to scramble the coupling oscillations, especially as their period decreases. Specific phenomena appear in the intermediate situation,  $D \sim \Delta$ : a higher order term is produced in the expansion of the interlayer coupling, known as the biquadratic coupling, which favours a relative orientation perpendicular to the magnetisations of the layers  $F_1$  and  $F_2$  [83]. This biquadratic term is often measured, but may have different origins.

Another problem commonly observed in multilayers is the problem of pinholes in the layer M, which allow direct contact and hence a strong local ferromagnetic coupling between layers  $F_1$  and  $F_2$ . This type of defect is shown in Fig. 14.46. It becomes critically important when one seeks to observe an antiferromagnetic interlayer interaction through the layer M. The ferromagnetic interaction through a pinhole, although it only affects a very small area, may have a strength per unit area that is greater by several orders of magnitude than the strength of the antiferromagnetic interaction, wherein the latter may be completely masked. An analogous argument to the one given above, taking into account the average distance between pinholes, can be used to make a quantitative estimate of their influence [84].

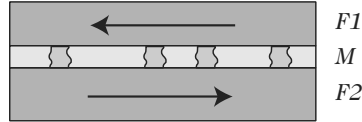
### Step-Edge Effects Due to Roughness

To calculate the anisotropy, a local quantity, the edges of the terraces must be taken into account. At atomic sites at the edge of a terrace, the atomic environment differs from that of a perfect surface. The presence of roughness thus modifies the interface magnetocrystalline anisotropy. If this roughness is oriented, e.g., parallel terraces in a vicinal surface, it can even induce a magnetic anisotropy in the plane of the film. The Néel pair model (see p. 518 [10]) can be used to analyse these effects, at least qualitatively [85]. Clearly, they will be more pronounced as  $D$  decreases and  $\delta e$  increases.

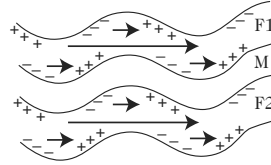
At the same time, on a rough film, the dipole anisotropy constant is lower than the value given by (14.26) [86]. Indeed, when the magnetisation is oriented in the plane of the film, there will be charges at the terrace edges which increase the minimal energy.

Finally, in a magnetic multilayer such as the  $F_1/M/F_2$  trilayer shown in Fig. 14.15, the main roughness, often due to the substrate, is generally conformal, i.e., at each interface, it reproduces the interface just below, as shown





**Fig. 14.46.** Pinholes in a magnetic layer



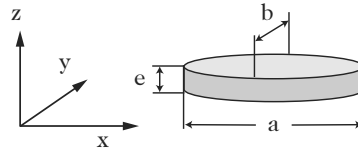
**Fig. 14.47.** Conformal roughness in a trilayer

in Fig. 14.47. This roughness induces magnetic charges on either side of the intercalating film M. Although they are weak charges, they can be very close to one another if the layer M is very thin. The corresponding dipole energy favours a ferromagnetic configuration, as shown in the figure, where magnetic charges of opposite sign are the closest. This effect, predicted by Néel [87], is known as orange peel coupling. It is generally a weak effect, with equivalent fields of the order of mT, but it can be a nuisance in magnetoelectronic devices. Recall that, in the absence of roughness, there is no dipole coupling between two ferromagnetic layers of infinite extent. Indeed, the field created by the magnetisation of the film vanishes outside the film, just as the electric field vanishes outside an infinite plane capacitor.

### Quantum Size Effects and Roughness

The above arguments apply perfectly to local quantities like the anisotropy. However, the interlayer oscillating coupling is not a local quantity, since quantum confinement effects require phase coherence of the electron wave function over distances at least as great as the thicknesses of the layers. When there are defects such as roughness, dislocations, interdiffusion, etc., capable of reducing this coherence, this has the effect of directly interfering with the oscillations.

The effect on the fringes is expressed through an exponential prefactor  $\exp(e_M/e_D)$  in (14.36), where  $e_D$  is the attenuation length due to decoherence. The theoretical study of  $e_D$  is no simple matter [39]. One may nevertheless come to grips with several important aspects by referring to the discussion in Sect. 14.1.4. Destroying the lateral structural coherence of interfaces  $F_1/M$  and  $M/F_2$  amounts in effect to allowing a small component  $k_{\parallel}$  in the expression for the vector  $q_{\perp}$  which determines the oscillation period. This effect can be understood by imagining, for example, that locally, due to the roughness, the interfaces are no longer strictly parallel with the average crystallographic direction of the film. Tilting the vector  $q_{\perp}$  changes its length, so that the defects introduce a distribution of periods which tend to scramble the oscillatory



**Fig. 14.48.** Thin elliptical wafer

pattern. The effect may depend strongly on the crystallographic direction. For example, for an intercalating layer of Cu (see Fig. 14.19), the variation of the length of  $q_{\perp}$  with its inclination is very small for  $q_2$  and  $q_1$ , which correspond to the two periods for the [001] direction. However, the variation is very large for the vector  $q_3$  along the [111] direction. The latter oscillation will thus be much more sensitive to the defects. In this case, the effect is correlated with the fact that the angle between the Fermi velocity of the electrons ‘carrying’ the quantum size effect and the [111] direction perpendicular to the film is very large (see Fig. 14.19). These electrons thus feel the lateral coherence of the crystal lattice more strongly.

It is difficult to corroborate this experimentally due to the difficulty in establishing the precise structure of the buried interfaces. It is worth mentioning two interesting cases:

- the study of oscillations in an Fe/Cr/Fe(001) trilayer as a function of the growth mode of the Cr layer [88],
- the direct study by inverse photoemission of the influence of the angle of inclination of a vicinal surface with perfectly controlled roughness on the oscillations of the electron density of states in a Cu layer on a Co(001) monocrystal [89].

The very rapid decrease ( $e_D$  of the order of 5 atomic planes) in the coupling oscillations in Au/Co/Au(111) trilayers can also be interpreted as a decoherence effect introduced by the existence of a dense network of interface dislocations between two metals in which the lattice interval differs by almost 15% [90].

### **B. Example of Complex Precession: Magnetisation Reversal in an Elliptical Wafer**

Let us return to the example of a thin elliptical wafer, as shown in Fig. 14.48, assumed to have a perfect ellipsoidal shape and to be small enough to be treated as a macrospin. We also assume a uniaxial magnetocrystalline anisotropy with axis  $Ox$ .

The demagnetising field  $\mathbf{H}_D$  is then given by (14.42) and the anisotropy field  $\mathbf{H}_A$  by (14.43). If the thickness  $e$  is very small compared with the lateral dimensions  $a$  and  $b$ ,  $N_X$  and  $N_Y$  are small compared with  $N_Z$ , which is thus of order 1 (see p. 512). Moreover, with typical metals,  $2K/\mu_0 M_S$  is always small compared with  $M_S$ . In these conditions, it can be shown that the anisotropy

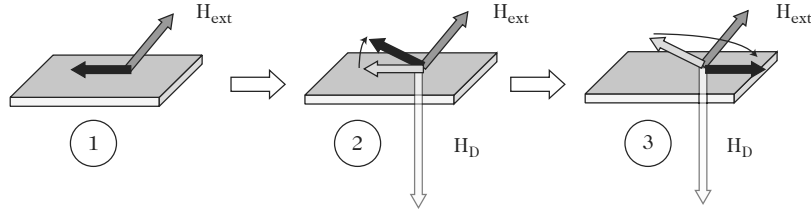


Fig. 14.49. Application of external field

of the wafer is correctly modelled by using  $N_Z = 1$  and integrating  $N_X$  and  $N_Y$  in a planar effective anisotropy field given by

$$\mathbf{H}_{A\text{eff}} = \left[ \frac{2K}{\mu_0 M_S} + (N_Y - N_X)M_S \right] m_X \mathbf{x} = \frac{2K_{\text{eff}}}{\mu_0 M_S} m_X \mathbf{x}.$$

When there is no damping, the dynamical equation becomes

$$\frac{d\mathbf{M}}{dt} = -\gamma (\mathbf{M} \times \mu_0 \mathbf{H}_{\text{eff}}),$$

where  $\mathbf{H}_{\text{eff}} = -M_S m_Z \mathbf{z} + \mathbf{H}_{A\text{eff}} + \mathbf{H}_{\text{ext}}$ , and  $\mathbf{H}_{\text{ext}}$  is the applied field.

Applying a magnetic field pulse with short enough rise time compared with the characteristic precession frequency of the system, the magnetisation will immediately trigger a precession dynamics. This dynamics can be used to reverse the magnetisation from the initial direction  $-\mathbf{x}$  to the new direction  $+\mathbf{x}$ .

The initial speed  $d\mathbf{M}/dt$  is maximised by first making  $\mathbf{H}_{\text{eff}}$  as close to perpendicular to  $\mathbf{M}$  as possible. This is done by applying an external field  $\mathbf{H}_{\text{ext}}$  perpendicular to  $\mathbf{M}$ . It is tempting to apply  $\mathbf{H}_{\text{ext}}$  along  $Oz$ , i.e., perpendicular to the wafer, which triggers a precession in the plane in conformity with the original aim. However, simulations show that this configuration requires high fields  $\mathbf{H}_{\text{ext}}$  in order to obtain a rapid rotation, and the trajectory is not simple because damping, neglected here, tends to pull the magnetisation out of the plane, and the field  $\mathbf{H}_{Dz}$  thereby produced opposes the rotation. It is in fact much more effective to apply  $\mathbf{H}_{\text{ext}}$  along  $Oy$ , as shown schematically in Fig. 14.49 [54].

As soon as the field is applied (phase 1), the couple produced by  $\mathbf{H}_{\text{ext}}$  tends to pull the magnetisation  $\mathbf{M}$  out of the plane (phase 2). The perpendicular component of  $\mathbf{M}$  immediately gives rise to a strong dipole field  $\mathbf{H}_{Dz}$ , perpendicular to the plane, whose couple triggers a precession of  $\mathbf{M}$  around  $Oz$  and towards the  $+\mathbf{x}$  direction, whilst at the same time bringing it back into the plane (phase 3). If the field  $\mathbf{H}_{\text{ext}}$  is then suppressed, an almost optimal magnetisation reversal has been obtained, with almost minimal energy expenditure. If  $\mathbf{H}_{\text{ext}}$  is left for longer, the symmetric configuration of phase 1 is retrieved. The couple exerted by  $\mathbf{H}_{\text{ext}}$  will pull  $\mathbf{M}$  out of the plane and into the direction opposite to that in phase 2. This gives a dipole field with the

opposite sign to the one in phase 2 and hence a precessional motion about  $Oz$  in the opposite direction to the one in phase 3, which brings the magnetisation into the  $-\mathbf{x}$  direction. When there is no damping, the motion continues in this way, following a global trajectory which corresponds to a ‘flattened’ precession about the direction of  $\mathbf{H}_{\text{ext}}$ : due to the dipole anisotropy, its extent out of the plane remains small compared with its amplitude in the plane.

A detailed analytical calculation of this trajectory can be found in [91], together with a discussion of the way it is influenced by energy dissipation until the magnetisation stabilises. From the point of view of applications, this type of precessional reversal has actually been observed in magnetoresistive devices with structures similar to those in Fig. 14.1 [92,93], with reversal times of the order of 150 ps for rise times of the field  $\mathbf{H}_{\text{ext}}$  of the order of 50 ps. The criterion for obtaining precession remains compatible with the performances of electronic circuits. Moreover, structures with sizes on the micron scale exhibit precessional behaviour very close to that of a nanoparticle and hence a perfectly reproducible reversal, in contrast to what is observed for the same structures in quasi-static fields.

## References

1. Baibich, M.N., Broto, J.M., Fert, A., Dau, F.V.N., Petroff, F., Etienne, P., Creuzet, G., Friederich, A., and Chazelas, J.: *Phys. Rev. Lett.* **61**, 2472 (1988)
2. Binash, G., et al.: *Phys. Rev. B* **39**, 4828 (1989)
3. du Tremolet de Lacheisserie, E. (Ed.): *Magnétisme*, Vol.1: *Fondements* and Vol.2: *Matériaux et Applications*, Presses Universitaires de Grenoble (1999)
4. Cohen-Tannoudji, C., Diu, B., and Laloë, F.: *Mécanique Quantique*, Hermann, Paris (1973)
5. MacIntosh, A.R., and Andersen, O.K.: In: *Electrons at the Fermi Surface*, ed. by M. Springford, Cambridge University Press (1980)
6. Ashcroft, N.W., and Mermin, N.D.: *Solid State Physics*, Saunders College (1976)
7. Bruno, P.: *Phys. Rev. B* **39**, 865 (1989)
8. Pauthenet, R.: *J. Appl. Phys.* **53**, 8157 (1982)
9. Bruno, P.: In: *24. IFF Ferienkurse des Forschungszentrum Jülich*, ISBN 3-89336-110-3 (1993)
10. Néel, L.: *Compt. Rend. Acad. Sci.* **237**, 1468 (1953); Néel, L.: *J. Phys. Rad.* **15**, 376 (1954)
11. Kittel, C.: *Phys. Rev.* **70**, 965 (1946)
12. Cowburn, R.P., et al.: *Phys. Rev. Lett.* **83**, 1042 (1999)
13. Shinjo, T., et al.: *Science* **289**, 930 (2000)
14. Cowburn, R.P., et al.: *Phys. Rev. Lett.* **72**, 2041 (1998)
15. Fruchart, O., et al.: *Phys. Rev. B* **63**, 174418 (2001)
16. Wernsdorfer, W., et al.: *Phys. Rev. Lett.* **77**, 1873 (1996)
17. Hubert, A., and Rave, W.: *Magnetic Domains*, Springer-Verlag (1998)
18. Rave, W., et al.: *J. Magn. Magn. Mater.* **190**, 332 (1998)
19. van den Berg, H.A.M.: *IBM J. Res. Dev.* **33**, 540 (1989)

20. Riedel, H., and Seeger, A.: *Phys. Stat. Sol.* **46**, 377 (1971)
21. Miltat, J.: In: *Applied Magnetism*, ed. by R. Gerber, C.D. Wright, and G. Asti, Kluwer Press (1994), pp. 221–308
22. McMichael, R.D., and Donahue, M.J.: *IEEE Trans. Magn.* **33**, 4167–4169 (1997)
23. Grolier, V.: Doctoral thesis, University of Paris Sud, 15 March 1994; Grolier, V., et al.: *J. Appl. Phys.* **73**, 5939 (1993)
24. Gradmann, U., and Müller, J.: *Phys. Stat. Sol.* **27**, 313 (1968)
25. Bland, J.A.C., and Heinrich, B. (Eds.): *Ultrathin Magnetic Structures I*, Springer-Verlag (1994) Chap. 2
26. McGee, N.W.E., et al.: *J. Appl. Phys.* **73**, 3418 (1993)
27. Ujfalussy, B., et al.: *Phys. Rev. Lett.* **77**, 1805 (1996). This theoretical paper cites several previous experimental references which it explains perfectly
28. Lee, C.H., et al.: *Phys. Rev. B* **42**, 11384 (1990)
29. Chappert, C., and Bruno, P.: *J. Appl. Phys.* **64**, 5737 (1988)
30. Jungblut, R., et al.: *J. Appl. Phys.* **75**, 6424 (1994)
31. Stöhr, J., and König, H.: *Phys. Rev. Lett.* **75**, 3748 (1995)
32. Weller, D., et al.: *Phys. Rev. Lett.* **75**, 3752 (1995)
33. Gambardella, P., et al.: *Nature* **416**, 301 (2002)
34. Gambardella, P., et al.: *Science* **300**, 1130 (2003)
35. Parkin, S.S.P., More, N., and Roche, K.: *Phys. Rev. Lett.* **64**, 2304 (1990)
36. Parkin, S.S.P.: *Phys. Rev. Lett.* **67**, 3598 (1991)
37. Bruno, P.: *Phys. Rev. B* **52**, 411 (1995)
38. Bruno, P., and Chappert, C.: *Phys. Rev. Lett.* **67**, 1602 (1991); *Phys. Rev. Lett.* **67**, 2592 (1991)
39. Bruno, P., and Chappert, C.: *Phys. Rev. B* **46**, 261 (1992)
40. Unguris, J., Celotta, R.J., and Pierce, D.T.: *J. Magn. Magn. Mater.* **127**, 205 (1993)
41. Unguris, J., Celotta, R.J., and Pierce, D.T.: *Phys. Rev. Lett.* **79**, 2734 (1997)
42. Garrison, K., Chang, Y., and Johnson, D.: *Phys. Rev. Lett.* **71**, 2801 (1993)
43. Carbone, C., Vescovo, E., Rader, O., Gudat, W., and Eberhardt, W.: *Phys. Rev. Lett.* **71**, 2805 (1993)
44. Okuno, S.N., and Inomata, K.: *Phys. Rev. Lett.* **70**, 1711 (1993)
45. Parkin, S.S.P., Chappert, C., and Herman, F.: *Europhys. Lett.* **24**, 71 (1993)
46. Unguris, J., Celotta, R.J., and Pierce, D.T.: *Phys. Rev. Lett.* **67**, 140 (1991)
47. Bloemen, P.J.H., et al.: *Phys. Rev. Lett.* **72**, 764 (1994)
48. Okuno, S.N., and Inomata, K.: *Phys. Rev. Lett.* **72**, 1553 (1994)
49. Johnson, M.T., et al.: *Phys. Rev. Lett.* **75**, 4686 (1995)
50. Majkrzak, C.F., et al.: *Phys. Rev. Lett.* **56**, 2700 (1986)
51. Geerts, W., et al.: *Phys. Rev. B* **50**, 12581 (1994)
52. Mégy, R., et al.: *Phys. Rev. B* **51**, 5586 (1995)
53. Bruno, P., Suzuki, Y., and Chappert, C.: *Phys. Rev. B* **53**, 9214 (1996)
54. Miltat, J.: An introduction to micromagnetics in the dynamic regime. In: *Spin Dynamics in Confined Magnetic Structures I*, ed. by B. Hillebrands and K. Ounadjela, Springer-Verlag (2002)
55. Landau, L.D., and Lifshitz, E.M.: *Phys. Zeit. Sowjetunion* **8**, 153–159 (1935)
56. Gilbert, T.L.: *Phys. Rev.* **100**, 1243 (1955)
57. Stoner, E.C., and Wohlfarth, E.P.: *Philos. Trans. R. Soc. London Ser. A* **240**, 599 (1948)

58. Thiaville, A.: Phys. Rev. B **61**, 12221 (2000)
59. Bonet, E.: Phys. Rev. Lett. **83**, 4188 (1999)
60. Jamet, M., et al.: Phys. Rev. Lett. **86**, 4676 (2001)
61. Pfeiffer, H.: Phys. Sta. Sol. **118**, 295 (1990)
62. Victora, R.H.: Phys. Rev. Lett. **63**, 457 (1989); Phys. Rev. Lett. **65**, 1171 (1990)
63. Thirion, C., Wernsdorfer, W., and Maily, D.: Nature Materials **2**, 524–527 (2003)
64. Néel, L.: Ann. Geophys. **5**, 99 (1949)
65. Brown, W.F.: Phys. Rev. **130**, 1677 (1963)
66. Grinstein, G., and Koch, R.H.: Phys. Rev. Lett. **90**, 207201 (2003)
67. Coffey, W.T., et al.: Phys. Rev. Lett. **80**, 5655 (1998)
68. Wernsdorfer, W., et al.: Phys. Rev. Lett. **78**, 1791 (1997)
69. Rizzo, N.D., Silva, T.J., Kos, A.B.: Phys. Rev. Lett. **83**, 4876 (2002)
70. Wernsdorfer, W., et al.: Phys. Rev. Lett. **79**, 4014 (1997)
71. Wernsdorfer, W.: Adv. Chem. Phys. **118**, 99–190 (2001)
72. Slonczewski, J.: J. Magn. Magn. Mater. **159**, L1–L7 (1996)
73. Berger, L.: Phys. Rev. B **54**, 9353–9358 (1996)
74. Stiles, M.D., Zangwill, A.: Phys. Rev. B **66**, 014407 (2002)
75. Albert, F.J., et al.: Appl. Phys. Lett. **77**, 3809 (2000)
76. Stiles, M.D., Zangwill, A.: J. Appl. Phys. **91**, 6812 (2003)
77. Kiselev, S.I.: Nature **425**, 380 (2003)
78. Osborn, J.A.: Phys. Rev. **67**, 351 (1945)
79. Aharoni, A.: J. Appl. Phys. **83**, 3432 (1998)
80. Cowburn, R.P.: J. Appl. Phys. **93**, 9310 (2003)
81. Bayreuther, G., et al.: J. Appl. Phys. **93**, 8230 (2003)
82. Dieny, B., and Vedyayev, A.: Europhys. Lett. **25**, 723 (1994)
83. Slonczewski, J.C.: Phys. Rev. Lett. **67**, 3172 (1991)
84. Bobo, J.F., et al.: Phys. Rev. B **60**, 4131–4141 (1999)
85. Bruno, P.: J. Phys. F: Met. Phys. **18**, 1291 (1988)
86. Bruno, P.: J. Appl. Phys. **64**, 3153 (1988)
87. Néel, L.: C.R. Acad. Sci. **255**, 1676 (1962)
88. Pierce, D.T., Stroschio, J.A., Unguris, J., and Celotta, R.J., Phys. Rev. B **49**, 14564 (1994)
89. Ortega, J.E., and Himpsel, F.: Appl. Phys. Lett. **64**, 121 (1994)
90. Grolier, V., et al.: Phys. Rev. Lett. **71**, 3023 (1993)
91. Devolder, T., and Chappert, C.: Eur. Phys. J. B **36**, 57 (2003)
92. Schumacher, A., et al.: Phys. Rev. Lett. **90**, 17201 (2003)
93. Schumacher, A., et al.: Phys. Rev. Lett. **90**, 17204 (2003)
94. de Vries, J.J., et al.: Phys. Rev. Lett. **75**, 4306 (1995)
95. Baibich, M., Broto, J.M., Fert, A., Guyen Van Dau, F.N., Petroff, F., Etienne, P., Creuzet, G., Friederich, A.: Phys. Rev. Lett. **61**, 2472 (1988)
96. Binash, G., Grünberg, P., Saurenbach, F., Zinn, W.: Phys. Rev. B **39**, 4828 (1989)
97. Schad, R., Potter, C.D., Beliën, P., Verbanck, G., Moshchalkov, V.V., Bruynseraede, Y.: Appl. Phys. Lett. **64**, 3500 (1994)
98. Dieny, B., Speriosu, V.S., Parkin, S.S.P., Gurney, B.A., Wilhoit, D.R., Mauri, D.: Phys. Rev. B **43**, 1297 (1991)
99. Pratt, W.P., Lee, S.F., Slaughter, J.M., Loloee, R., Schroeder, P.A., Bass, J.: Phys. Rev. Lett. **66**, 3060 (1991)

100. Piraux, L., et al.: Appl. Phys. Lett. **65**, 2484 (1994); Blondel, A., et al.: Appl. Phys. Lett. **65**, 3019 (1994)
101. Fert, A., Piraux, L.: Special issue: *Magnetism Beyond 2000*, J. Magn. Magn. Mat. **200**, 338 (1999)
102. Gijs, M.A.M., Johnson, M.T., Reinders, A., Huisman, P.E., van de Veerdonk, R.J.M., Lenczowski, S.K.J., Gansewinkel, R.M.J.: Appl. Phys. Lett. **66**, 1839 (1995)
103. Parkin, S.S.P., More, N., Roche, K.R.: Phys. Rev. Lett. **64**, 2304 (1990)
104. Slonczewski, J.: J. Magn. Magn. Mat. **159**, 1 (1996)
105. Katine, A., et al.: Phys. Rev. Lett. **84**, 3149 (2000); Albert, F.J., Katine, J.A., Burhman, R.A., Ralph, D.C.: Appl. Phys. Lett. **77**, 3809 (2000)
106. Grollier, J., Cros, V., Hamzic, A., George, J.M., Jaffrès, H., Fert, A., Faini, G., Ben Youssef, J., Legell, H.: Appl. Phys. Lett. **78**, 3663 (2001)
107. Moodera, J., Kinder, L.R., Wong, R.M., Meservey, R.: Phys. Rev. Lett. **74**, 3273 (1995)
108. Sommerfeld, A., Bethe, H.: *Handbuch der Physik*, Verlag Julius Springer, Berlin (1933) Vol. 24, p. 333
109. Jullière, M.: Phys. Lett. **54** A, 225 (1975)
110. Bowen, M., Bibes, M., Barthélémy, A., Contour, J.P., Anane, A., Lemaître, Y., Fert, A.: Appl. Phys. Lett. **82**, 233 (2003)
111. De Teresa, J.M., Barthélémy, A., Fert, A., Contour, J.P., Montaigne, F., Senneor, P.: Science **286**, 507 (1999)
112. Oleinik, I.I., Tsymbal, E.Y., Pettifor, D.G.: Phys. Rev. B **62**, 3952 (2000); Oleinik, I.I., Tsymbal, E.Y., Pettifor, D.G.: submitted to Phys. Rev. B Rapid Comm.
113. Kikkawa, J.M., Aschwalom, D.D., Smorchkova, I.P., Samarth, N.: Science **277**, 1284 (1997)
114. Gardelis, S., et al.: Phys. Rev. B **60**, 7764 (1999); Hammar, P.R., et al.: Phys. Rev. Lett. **83**, 203 (1999); Monzon, F.G., et al.: Phys. Rev. Lett. **84**, 5022 (2000); Zhu, H.J., et al.: Phys. Rev. Lett. **87**, 016601 (2001)
115. However, it should be noted that a polarisation of 12% of the light emitted by a spin LED was measured by Hanbicki, A.T., et al.: Appl. Phys. Lett. **80**, 1240 (2002). This proves the efficiency of injection from Fe into AlGaAs. This result has been attributed to the existence of a narrow Schottky barrier at the Fe/AlGaAs interface
116. Fiederling, R., et al.: Nature **402**, 787 (1999)
117. Park, Y.D., et al.: Appl. Phys. Lett. **77**, 3989 (2000)
118. R. Mattana et al.: Phys. Rev. Lett. **90**, 166601 (2003)
119. Datta, S., and Das, B.: Appl. Phys. Lett. **56**, 665 (1990)
120. Dietl, T., et al.: Science **287**, 1019 (2000)
121. Mott, N.: Proc. Roy. Soc. **156**, 368 (1936)
122. Campbell, I.A., and Fert, A.: *Ferromagnetic Materials*, ed. by E.P. Wohlfarth, North Holland, Amsterdam (1980) p. 769
123. Mertig, I., Zeller, R., and Diederich, P.H.: Phys. Rev. B **47**, 16178 (1993)
124. Fert, A., and Campbell, I.A.: J. Phys. F **6**, 849 (1976)
125. Vouille, C.: Barthélémy, A., Elokani, F., Fert, A., Schroeder, P.A., Hsu, S.Y., Reilly, A., Loloee, R., Pratt, W.P.: Phys. Rev. B **60**, 6710 (1999)
126. Camley, R.E., and Barnas, J.: Phys. Rev. Lett. **63**, 664 (1989)
127. Johnson, B.L., and Camley, R.E.: Phys. Rev. B **44**, 9997 (1991)

128. Barnas, J., Fuss, A., Camley, R.E., Grünberg, P., and Zinn, W.: Phys. Rev. B **42**, 8110 (1990)
129. Sondheimer, E.H.: Adv. Phys. **1**, 1 (1952)
130. Bardeen, J.: Phys. Rev. Lett. **6**, 57 (1961)
131. Wolf, E.L.: *Principles of Tunneling Spectroscopy*, Clarendon Press, Oxford (1985)
132. Moodera, J., and Mathon, G.: J. Magn. Magn. Mat. **200** (1999)



## Information Storage

D. Fraboulet and Y. Samson

When it comes to archiving, storing, and recording information, either temporarily or permanently, the needs of modern society never cease to grow. It is not easy to draw up a panoramic view of this phenomenon, given the enormous range of storage media and devices that have been devised to answer so many specific data storage requirements. However, a basic division can be made between permanent storage of large amounts of data (hard disks, CDs, DVDs, etc.) and storage associated with data processing (active and read only memory in computers, flash compact cards in digital cameras, etc.). Table 15.1 illustrates some of these features.

Whilst the cornerstone of microelectronics is the logic system, information storage is no less important. Indeed, the proportion of storage circuits is on the increase in mass-produced microchips (from 10 to 15%). These storage elements share a matrix architecture, in which reading and writing are carried out solely by electrical programming and in a solid phase resulting from an integration process of microelectronic type.

### 15.1 Mass Memories

In CDs and DVDs, information is read, and written in the case of writable disks, by a laser. The writing process is a thermal one in which local heating deforms a polymer layer (CD, CD-R), or induces local amorphisation of a phase-changing material such as a GeSbTe alloy (CD-RW, DVD). Reading is always based on detection of a contrast in optical reflectivity between different regions of the disk. We shall not go into further description of these systems, but rather focus on the hard disk and the microprobe memories which may supplant it, and matrix memories. Indeed, continuous progress achieved over several decades has carried this kind of technology to the gates of the nanoworld. Of course, this term applies naturally given the characteristic sizes of these components, but also because the very operation of such systems has

**Table 15.1.** Comparison of high-performance products available in 2003. Capacities are given in bits, although commercially they are more commonly expressed in bytes. Densities are given in bit/cm<sup>2</sup>. One often finds a unit of bit/inch<sup>2</sup>, where 1 inch = 2.54 cm. The access time is the time required for the read head (laser spot) to move to the region in the medium where data is stored. In a hard disk, this includes the time taken to transfer the read head to the region containing the required piece of information. The retention time is the time after which the information is lost if not refreshed. In matrix memories, writing and reading are based upon very different physical processes which may require very different times

Mass memories	Technology	Capacity	Density [Gbit/cm <sup>2</sup> ]	Access time	Optimal data flow <sup>a</sup>	Retention time	Information stable with power off
	Hard disk	800 Gbit	3.5	Average 7 ms	700 Mbit s <sup>-1</sup>	> 10 yr	Yes
	CD	5.6 Gbit	0.05	80 ms	62 Mbit s <sup>-1</sup>	> 10 yr	Yes
	DVD	37.6 Gbit (single disk)	0.4	80 ms	160 Mbit s <sup>-1</sup>	> 10 yr	Yes
	Microprobe	Tbit	60–200	?	~ 1 kbit s <sup>-1</sup>	Depends on medium	Yes
Matrix memories	Technology	Capacity	Density <sup>c</sup> [Gbit/cm <sup>2</sup> ]	Write time	Read time	Retention time	Information stable with power off
	DRAM	1 Gbit	1	10 ns	10 ns	70 ms	No
	SRAM	20 Mbit	0.1	~ 1 ns	~ 1 ns	> 10 yr	No
	Flash	256 Mbit	0.5	20 ns	~ 1 μs	> 10 yr	Yes
	FRAM	64 Mbit	0.2	10–50 ns	< 100 ns	> 10 yr	Yes
	MRAM	4 Mbit (demonstrator) <sup>e</sup>	0.5–1 <sup>d</sup>	15 ns	< 25 ns	> 10 yr	Yes
	PCRAM	<sup>e</sup>	0.5–1 <sup>d</sup>	15 ns	< 100 ns	> 10 yr	Yes

<sup>a</sup> This refers to sequential data readout once the read head has been positioned over a given region of the medium.  
<sup>b</sup> Difficult to establish accurately. The flow depends on the number of microprobes operating in parallel to read and write data (from a few 100 to a few 1000). For mechanical reasons, architectures currently envisaged restrict the flow rate per probe to < 10 kbit/s.  
<sup>c</sup> Density of zone containing memory cells. Today's memories include a peripheral zone (connections, etc.) with approximately equivalent area, leading to a true density of about half the value.  
<sup>d</sup> These figures are loose estimates for concepts still under development.  
<sup>e</sup> Capacities comparable with those of the DRAM can be envisaged with this type of memory.

recently come up against physical and technological problems that are specifically associated with nanometric dimensions. The aim here will be to describe the technology used today and in particular its limitations, and then discuss current ideas for pursuing the road to very high densities.

### 15.1.1 Mass Memories: The Hard Disk

The hard disk undoubtedly holds an important place amongst mass-produced data storage tools. With its extraordinarily rapid evolution over the last few decades, its low cost and relative reliability as a carrier of erasable and rewritable information, it now plays a part in our everyday lives, both at home and at work. Today, it has such a large capacity that it represents a real challenge to optical media in new products such as digital video recorders. Over the last 20 years, this extraordinary progress has carried this technology from the micro- to the nanoworld. Remarkably, the hurdles encountered in the pursuit of technological progress have meant that recording densities have advanced in leaps and bounds: 25% per year up to 1992, then 60% when IBM introduced magnetoresistive read heads, and 100% or more when IBM introduced giant magnetoresistive read heads in 1999 (see below). Recently, with the advent of nanoscale bits, real changes have also become essential in the storage media used. We have grown used to this kind of progress in data storage capacity, which has so rapidly transformed our lives. However, we will not be able to pursue this development at the same rate and over such a long period unless we find some way of overcoming the specific obstacles that face us on the nanoscale.

#### The Magnetic Hard Disk: Principles and Problems

Up to 2005, all hard disks fall into the category of longitudinal storage media, i.e., having planar magnetic anisotropy. Data is held on a thin magnetic layer and coded by transitions between domains of opposite magnetisation. Storage media capable of densities upwards of 15 Gbit/cm<sup>2</sup> were demonstrated in 2002 [1]. Technically, two quantities are essential here: the width of the track followed by the read head, often expressed through the number of tracks per unit length, and the minimal distance between transitions along a track, often expressed in bits per unit length (kbit/cm). At the highest densities achieved so far, these two distances are of the order of 200 for the track width and 40 nm for the bit length.

#### Giant Magnetoresistive Read Head

Figure 15.1 shows the read head above a track on a medium with plane magnetisation (longitudinal medium). The write element consists of a small electromagnet brought very close to the surface of the medium (about 10 nm). When a current

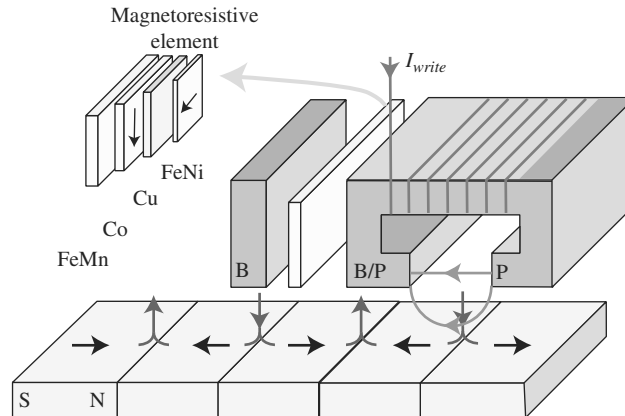


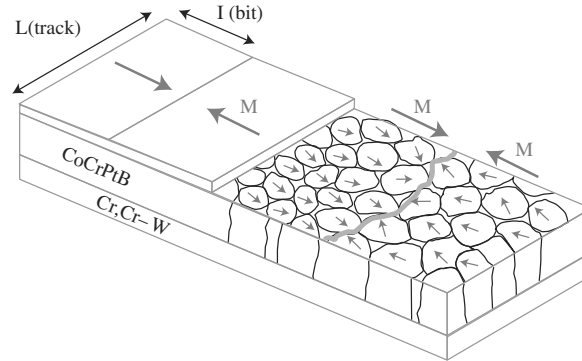
Fig. 15.1. GMR read head

pulse is applied through the microcoil, the magnetic material is polarised and produces a leakage field in the air gap (poles P) and in the medium. One of the poles also stands in as one of the two shield plates insulating the read element from perturbations associated with the leakage field of adjacent transitions. Reading is achieved by a spin valve: the resistance of the Co/Cu/FeNi trilayer depends on the respective orientations of the magnetisations in the two magnetic layers. The direction of magnetisation of the hard layer (cobalt-based) is pinned by exchange with an antiferromagnetic layer. The direction of magnetisation of the soft layer (FeNi) fluctuates readily under the effects of the leakage field from the medium.

- Writing is achieved by applying a current pulse in a coil, which polarises a magnetic material with strong magnetisation. A magnetic field is then produced locally inside and under the air gap, thereby inscribing a bit of data in the medium. The bulk magnetisation of the material making up the poles restricts the field that can be created in this way (to at most 2.4 T for currently available materials).
- Since 1999, the read element has exploited the phenomenon of giant magnetoresistance. The magnetoresistive element, comprising two magnetic layers and a non-magnetic spacer, is called a spin valve. Under the effect of the leakage field localised above the transitions of the magnetic medium, the magnetisation of the soft layer rotates, whilst the magnetisation of the reference layer remains stable. This leads to a change in the electrical resistance which can be detected.

The active layer of the medium is a thin layer (about 15 nm thick) made from small crystal grains (about 8 nm in size) of a cobalt alloy. Its design brings together a large number of technological tricks of the trade (see Fig. 15.2).

To compensate for statistical scatter in grain properties such as size, anisotropy axis, etc., and maintain an acceptable signal-to-noise ratio, a bit must be made from a large number of grains. Hence, sizes of 8–10 nm reached by the grains have led to the appearance of a new phenomenon connected with the nanoscale dimensions of these objects. This is superparamagnetism,



**Fig. 15.2.** Planar magnetic medium. Data is carried by magnetic transitions separating regions of opposite magnetisation. A strong magnetic field is located vertically above them. Each grain has an anisotropy axis oriented in the plane of the thin layer. This plane orientation of the anisotropy axes is obtained by epitaxial texturing with a chromium-based sublayer. The grain diameter is 7–8 nm in currently used storage media and there are several hundred grains per bit. In order to broaden the magnetic transitions between bits, the magnetisation of each grain must be able to behave relatively independently from its neighbours. This is achieved by forming non-magnetic precipitations at the grain boundaries (Cr, Pt, Ta, and more recently, B are commonly added to the cobalt), so as to reduce the exchange coupling between neighbouring grains. The dipolar coupling of the magnetic field between grains also has a detrimental effect, but it is unavoidable

wherein the magnetisation direction of each grain, and hence the information recorded, are no longer stable due to the simple fact of thermal fluctuations. (At the sizes in question here, it is realistic to assume that each grain has uniform magnetisation at any given time.) This problem has been known for several years now [2] and has become a major concern in the field of magnetic recording. Theoretically, a 10-year data lifetime corresponds to the condition  $K_u V / k_B T > 40$ , where  $V$  is the grain volume,  $K_u$  is the anisotropy, and  $k_B$  is Boltzmann's constant. This critical value increases rapidly for smaller grain sizes due to the growth of the demagnetising field for small dimensions. The switching time, i.e., the time elapsed between two magnetisation reversals, is described by an Arrhenius law:

$$\tau^{-1} = f_0 \exp \left[ -\frac{E(H)}{k_B T} \right], \quad (15.1)$$

where  $E(H)$  is the barrier energy that has to be overcome in order to reverse the magnetisation of the particle, and  $f_0$  is the attempt frequency, whose value for the kind of materials considered here is commonly taken to be around  $f_0 \sim 10^9$  Hz.

The three terms in (15.1) correspond to as many solutions to the problem of paramagnetism. The first involves acting on the temperature, by making

**Table 15.2.** Properties (anisotropy  $K_u$  and coercive field  $H_c$ ) and superparamagnetic limiting diameter  $D_c = 60k_B T / K_u^{1/3}$ , for a stability of 10 years, for typical materials. Alloys FePd and FePt are considered as good candidates for making storage media capable of very high densities.  $M_s$  is the saturation magnetisation of the material. The critical diameter corresponds to a stable magnetisation direction for an isolated particle, defined here by  $K_u V / k_B T > 60$ . Note that the coercive field associated with very high anisotropies in materials like FePt is much higher than fields that can be achieved in practice by a small read head [2]

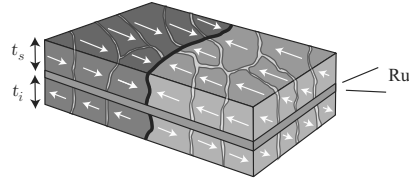
	Critical diameter [nm]			Magnetocrystalline anisotropy of material	Coercive field at 300 K ( $H_c = 2K_u/M_s$ )
	At 300 K	At 77 K <sup>b</sup>	At 4 K <sup>c</sup>		
Cobalt	8	5.1	1.9	0.45	0.64
FePd <sup>a</sup>	5	3.5	1.3	1.8	3.3
FePt <sup>a</sup>	3.5	2.2	0.8	7	11.6

<sup>a</sup> L1<sub>0</sub> structure. <sup>b</sup> Liquid N<sub>2</sub>. <sup>c</sup> Liquid He.

the hard disk operate below room temperature. For practical reasons, this is not considered to be a real option at the present time. Increasing the volume of the grains would work against attempts to increase the density, at least if the geometry of the medium remained unchanged. However, we shall see later that this idea has motivated a significant amount of research into discrete storage media (quantum disk media), and that it has given rise to the highly ingenious technology of antiferromagnetically coupled media. In the longer term, and for reasons described below, further progress should be possible using storage media with perpendicular magnetisation. Finally, materials with higher anisotropy can be used. Such things are available, e.g., chemically ordered FePt alloys. This channel has been used but it involves one serious disadvantage: the magnetic field  $H_c$  that the read head must produce to reverse the magnetisation of a grain grows rapidly with the anisotropy. (For a particle,  $H_c \propto K_u/M_s$ , where  $M_s$  is the saturation magnetisation of the material.)  $H_c$  will very quickly exceed any practically achievable values in the nanometric air gap of a read head. This solution thus leads to a new problem for which a novel solution has been devised: thermally assisted recording. This would open the way to using materials with very high anisotropies, such as L1<sub>0</sub> alloys (FePt, etc.). This technology exploits the fact that  $H_c$  decreases with temperature, because thermal agitation helps in overcoming energy barriers. We shall describe these various technologies.

### Antiferromagnetically Coupling Media

The problem raised is relatively complex and we shall only outline the main features here. For a medium characterised by grain diameter  $D$ , transitions



**Fig. 15.3.** Antiferromagnetically coupled medium. The medium consists of two magnetic layers separated by a thin layer of ruthenium, a non-magnetic metal. The thickness of the ruthenium layer is chosen to induce antiferromagnetic coupling between the two ferromagnetic layers. During writing the field from the head saturates the magnetisation of the two FM layers. The exchange field induced by the ruthenium layer is then strong enough to bring the magnetisation of the lower layer to an antiparallel direction

between bits, which follow the grain boundaries, cannot have widths less than about  $D/2$ . A great deal of effort must therefore be invested in reducing  $D$ . Moreover, the leakage field induced by the transition between bits and read by the recording head spreads out with height  $h$  above the medium. The width of the transition thus increases with the thickness  $t$  of the data-bearing layer and the magnetisation (a stronger demagnetising field broadening the magnetic transition between bits). This situation has led to a constant reduction in the thickness of the medium in the quest for higher densities. Of course, the amplitude of the signal also diminishes and this problem has only been solved by successive technological breakthroughs in the design of read elements, which must be ever more sensitive. But coming back to the question of the media themselves, the relevant term in the end is  $M_r t$ , where  $M_r$  is the remanent magnetisation of the magnetic layer.

The problem is to reduce  $t$  without simultaneously reducing the volume of the grains constituting the thin layer below the superparamagnetic limit. Concerning this apparently insoluble equation, a very elegant solution has been devised, which exploits the idea of media with antiferromagnetic coupling (see Fig. 15.3). These media consist of two separated ferromagnetic layers with antiparallel magnetisations. The antiferromagnetic coupling between the two ferromagnetic layers is obtained via a layer of ruthenium. It results simultaneously from the so-called RKKY coupling and coupling by the dipole field. For such a structure, the effective thickness  $t_{\text{eff}}$  is given by

$$M_r t_{\text{eff}} = M_r t_s - M_r t_i ,$$

where  $t_s$  and  $t_i$  are the respective thicknesses of the two magnetic layers. The advantage with these new media is two-fold [3]:

- The width of the transition pulse between bits is reduced because it is proportional to  $M_r t_{\text{eff}} = M_r (t_s - t_i)$ . This makes it possible to increase recording densities.

- The barrier energy  $KV_{\text{eff}}$  opposing magnetisation reversal due to thermal fluctuations is not proportionally reduced:  $KV_s < KV_{\text{eff}} < KV_s + KV_i$ . The magnetisation of grains carrying the recorded information is therefore stabilised. (The RKKY coupling corresponds to a surface coupling energy, and the last term of the equation would be obtained for perfect coupling.)

These media are now used in commercial hard disks and the improved data stability has pushed back the inevitable change of technology required as we approach the superparamagnetic limit. Their precise potential is not yet fully understood, mainly owing to the complexity of the phenomena associated with coupling between two ferromagnetic layers. The lower ferromagnetic layer is very thick and would probably be superparamagnetic without exchange with the upper layer.

### Perpendicularly Magnetised Media

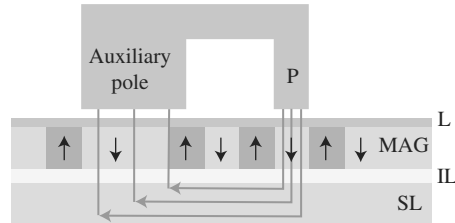
The idea of using layers with perpendicular magnetisation is not a new one and has had fervent supporters for almost 25 years now [4]. However, the extraordinarily rapid progress of planar media succeeded in postponing its implementation up to 2005<sup>1</sup> (first commercial disk with perpendicular magnetisation). The main arguments in favour of media with perpendicular magnetisation are as follows:

- New head geometries become possible, in which the bit is directly opposite a pole (see Fig. 15.4). Note the technological trick represented by the soft underlayer and the asymmetry between the poles, which allows the field to close near the read head with strengths below the coercive field of the medium. This new geometry also allows the head to produce higher fields in the medium, so that materials with higher coercivities, and correspondingly high anisotropies, can then be used. The medium can then be designed with smaller grains without threatening thermal stability, and thereby carry higher densities. Moreover, the suggested head geometry can produce a field in the medium that is more uniform with respect to distance from the pole, and this means that it will be possible to use thicker media, with correspondingly greater thermal stability.

---

<sup>1</sup> Beyond any simple statement of technical and scientific merits at a given moment of time, this is a key point for understanding the way information technology has evolved over the last few decades. Indeed, established technologies move forward very quickly, on the basis of huge investments. Hence, if a new technology is to impose itself, it must not only be more competitive at the moment it comes into being; it must also remain so in the face of the constant progress being made by established solutions during the inevitable delay required for industrialisation. Moreover, the advantages must be great enough to justify the corresponding financial and technological risks. This means that, in the specific sector of information technology, radically new technologies are rarely introduced until the fundamental limits of previous technologies become a real obstacle.





**Fig. 15.4.** Perpendicular medium and head. The medium (MAG) consists of a layer with perpendicular magnetisation, such as a Pt/Co multilayer with interface anisotropy or  $L1_0$  alloys like FePd, and a soft underlayer (SUL), separated from the active layer by a non-magnetic interlayer (IL). The soft underlayer channels the field lines, under the information-bearing layer, towards the auxiliary pole. The size ratio of the two poles is arranged to ensure a strong field under the main pole for writing purposes, and a weaker field (less than the coercive field of the medium) under the auxiliary pole. A thin carbon-based layer fulfills tribological requirements

- In contrast to what is observed for planar media, a reduction in bit size is favourable with regard to demagnetising field terms: field lines connect between neighbouring bits with opposite magnetisation, causing a mutual stabilising effect.
- It is relatively easy to obtain a perpendicular uniaxial anisotropy during epitaxial growth. By avoiding the statistical scatter of anisotropy axes observed in planar media, transitions between bits are cleaner and less noisy, whence higher densities can be achieved.

At the present time, a great deal of research is focused on making perpendicularly magnetised media with suitable performance. There are essentially two options: granular media based on CoCr(Pt), similar to those used in current media but in which the epitaxial relation creates a perpendicular anisotropy axis, and Co/Pt (or Co/Pd) multilayers. In the latter, the anisotropy obtained is an interface anisotropy, arising from the preferential spin orientation perpendicular to the Co/Pt interface.

### Heat-Assisted Recording

The problem here is relatively simple: by imposing a reduction in grain size, increased recording densities require the use of materials with higher anisotropy to counter the threat of superparamagnetism. Now the coercive field of a magnetic particle grows with its anisotropy ( $H_c \propto K_u/M_s$ ) and can easily reach fields greater than those that can be produced by the read head (about 2.4 T using materials with the highest magnetisation). The idea is to heat the medium locally during writing, taking advantage of the fact that the coercive field falls when the temperature is increased. The technology that comes closest to a practical application makes joint use of a laser to induce local heating

in the medium and a conventional read head to read and write the information. Despite the greater complexity of the read/write head, the advent of thermally assisted recording seems certain over the next few years.

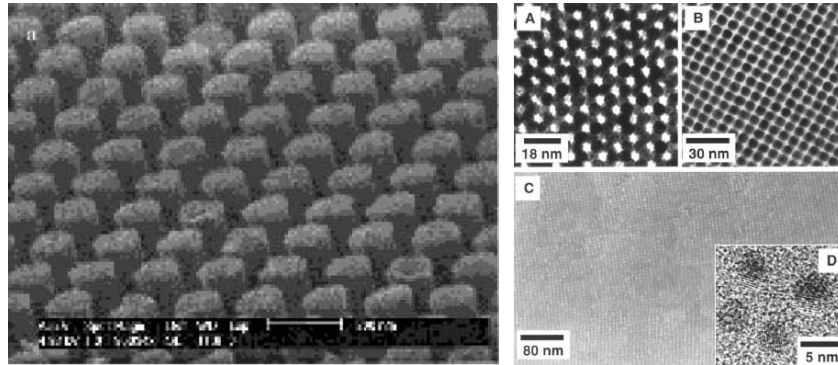
### Discrete Media

In currently used media, the size of magnetic grains carrying the data must be much less than the dimensions of the information bit so as to compensate, by an averaging effect, for the fluctuations in their properties (diameter, anisotropy axis, etc.). Moreover, since the grains are small enough to exhibit approximately uniform magnetisation, it is essential to reduce their size in order to be able to reverse the magnetisation direction over a short distance in the transition between bits, despite the residual coupling between grains. The simple setup illustrated in Fig. 15.5 shows the advantage of using a discontinuous medium: exchange coupling between dots is zero, and if each dot, consisting of a single grain, can carry one bit of information, the grain size here can be much higher for the same density. The performance of current media shows that the lateral dimensions of a dot will soon have to be less than 50 nm to achieve competitive densities. We shall now review some of the possibilities for doing this.

Optical lithographic techniques, limited by available wavelengths, cannot easily achieve such small dimensions. Electron lithography provides a credible alternative, often used to make demonstrators including nanoscale components. However, it is a slow and costly technique, rather ill-adapted to the fabrication of large numbers of identical nanometric objects. In this context, a new technique has been proposed wherein the pattern obtained on a first array is transferred a great many times by stamping (see Fig. 15.5, left) [5].

A significant disadvantage here is the fact that the medium loses planarity, an unacceptable situation for the read head, which must somehow be compensated. An alternative solution was thus envisaged in 1999 by Chappert et al. in 1999 [7], wherein the medium is structured laterally by modifying its magnetic properties locally but without removing any matter. The idea was applied to a magnetic multilayer (Pt/Co), where the interface anisotropy at the origin of the perpendicular magnetisation can be destroyed by the chemical mixing induced by irradiating with light helium ions (130 keV). At higher doses, the magnetic element dissolves in the platinum, making the medium locally non-ferromagnetic. This allows one to mark out the magnetic dots. A resolution of 10 nm is achievable by this means.

In direct competition with these techniques, work is being done to obtain magnetic dots directly by a self-organisation process. A good example is provided by the chemical synthesis and spontaneous organisation of 6-nm FePt particles after their growth on a platinum substrate (see Fig. 15.5, right) [6]. However, although self-organisation techniques constitute one of the favoured routes towards the nanoworld, and although they can indeed fabricate large numbers of objects with sizes quite inaccessible to the tools of lithography, it

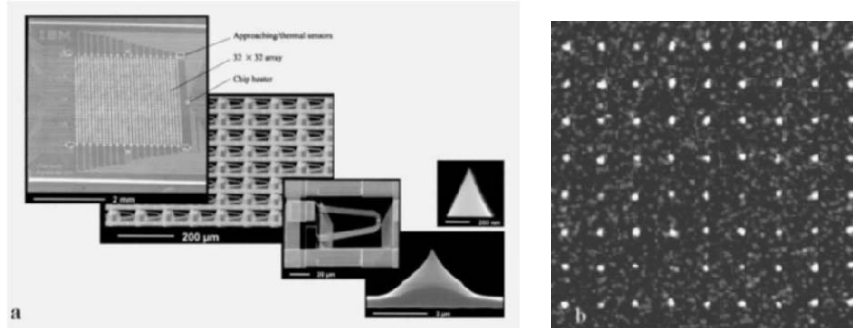


**Fig. 15.5.** Synthesising nanoscale magnetic objects. Advanced lithography and self-organisation. *Left:* Square array of nickel pillars, each 70 nm in diameter and 400 nm high, obtained by nanoimprinting. Schematically, a master array or template is obtained by electron lithography, a high resolution technique, but slow and hence costly. The pattern is then transferred by stamping the resist layer deposited during the first stage of the lithography onto another substrate. Image courtesy of W. Wu et al. [5]. *Right:* Electron microscope images of self-assemblies of FePt particles (diameter 6 nm) on a silicon oxide substrate [6]. The particles are obtained by reduction of platinum acetylacetonate  $\text{Pt}(\text{acac})_2$  and decomposition of  $\text{Fe}(\text{CO})_5$  in solution. The size of the particles can be controlled in the range 3–10 nm, with dispersion less than 5%. The spacing, 4 nm in (A) or 1 nm in (B), and the arrangement of particles, hexagonal in (A) or cubic in (B), depends on the ligands used. (C) Large scale arrangement of particles, showing a certain number of imperfections. Image courtesy of S. Sun et al. [6]. (D) High resolution electron microscope image of several nanoparticles

must be said that a perfect large scale organisation has so far proved elusive. Will a solution ever be found to counter this difficulty? From the answer to this question will emerge a large part of the future of nanotechnology. Generically, two outcomes may be considered: either it will become possible to eliminate the defects induced by the still imperfect self-organisation processes, or it will prove possible to design read and write processes able to tolerate these defects.

### 15.1.2 Beyond the Hard Disk. Local Probe Techniques

With the rapid increase in densities, it will doubtless soon become illusory to incorporate the complex read and write functions contained in today's read heads into a very small probe. It is thus clear that much of the technology with which we are now familiar will no longer be suitable for the age of the nanoworld and will have to be replaced. However, in order to take the upper hand, a new technology must manifest a considerable potential for further progress towards smaller dimensions. The investment required for a new technology is so great that it could not be made for a technique that would only provide a temporary application before being swept away by some



**Fig. 15.6.** Ultrahigh density microprobe recording. **(a)** The Millipede storage chip, developed by IBM Zurich, for the parallel operation of a large number of thermal sensor tips. *From left to right:* Array of  $32 \times 32$  cantilevers, each carrying a probe, used to inscribe and read information on a polymer medium; the second and third images show the details of each element, carrying a microprobe with nanoscale radius of curvature at its end (*fourth and fifth images*) [8]. **(b)** Memory dots. Writing achieved using conducting microprobes in a medium including a thin layer of the alloy  $\text{Ge}_2\text{Sb}_2\text{Te}_5$ , a phase-change material. The layer is amorphous and in a state of high electrical resistance after deposition. The image (900 nm) is a map of the local resistivity obtained by an atomic force microscope. Bright regions correspond to low electrical resistivity. These are the crystalline islands obtained by heating the medium locally by injecting a current pulse from a conducting microprobe. The diameter of the conducting islands is just  $12 \pm 3$  nm and their spacing is about 100 nm. Image courtesy of O. Bichet, O. Lemmonier, S. Gidon, Y. Samson, CEA Grenoble

other solution. With this in mind, an IBM research team suggested using simple silicon probes, similar to the tips used in atomic force microscopes [8]. Such probes can currently be made with a radius of curvature of a few nanometers. Research is therefore directed towards designing an appropriate medium, one which allows simple read and write processes and can carry very high densities. Moreover, for essentially mechanical reasons arising from the cantilever vibration frequencies (the cantilever is the name given to the little lever carrying the microprobe), and also the need to maintain a permanent contact between probe and surface for most writing processes so far envisaged, the probe cannot move as fast as the read head of a hard disk. To guarantee a high data flow rate, a great many read and write operations must be carried out simultaneously. Several research teams have thus been working on the design of probe arrays capable of operating in parallel (see Fig. 15.6a).

The Millipede is designed to read and write using heat transfer processes. During writing, a resistive element located vertically above the microprobe is heated to about  $200^\circ\text{C}$  by passing an electric current. The polymer in contact with the probe softens and the medium can be indented, leaving a depression several tens of nanometers across. Densities up to  $0.2 \text{ Tbit}/\text{cm}^2$  can

be achieved. Reading exploits the temperature dependence of the resistance of the heating element. The temperature of this element decreases when the probe is located in a depression inscribed on the medium, because a decrease in the pyramid–surface distance leads to higher thermal transfer.

Other media are currently being investigated for microprobe recording. Among these, it is worth mentioning thin magnetic layers and phase-change materials. The former have the advantage of being based on a considerable knowhow accumulated in the design of hard disks. However, they also suffer from the same limitations, such as the superparamagnetic limit. Moreover, it would seem that the same physical process cannot be used for the write and read operations on such media. This implies that a single probe could not be used to achieve these operations. The phase-change materials are generally chalcogenides (tellurium alloys such as GeSbTe). At room temperature, the amorphous and crystalline phases are both kinetically stable. This fact is exploited in optical disks such as rewritable CDs and DVDs, where the phase transition is induced locally by laser heating, at a moderate temperature to obtain crystallisation, and up to melting point (liquid phase) followed by rapid quenching to obtain amorphisation. The read operation involves detecting the contrast in optical reflectivity using the same laser. Several teams, in Japan and in Grenoble (France), have suggested using local heating caused by the Joule effect when a current is injected from the probe for writing, and the contrast in electrical resistivity between the two phases for reading (see Fig. 15.6b). Densities of the order of 1 Tbit/cm<sup>2</sup> have been demonstrated.

## 15.2 Matrix Memories

This heading is intended to cover memory circuits addressed only electrically and, in contrast to the mass memories discussed above, involving no moving parts. The aim will be to draw up a succinct picture of the state of the art in this field, both industrially (commercially available products) and with regard to medium and long term research. However, we have only included those ideas that have already proven themselves in a realistic matrix environment. Indeed, we shall see that it is not sufficient, and by a long way, to simply reproduce a single memory cell able to store a 0 or a 1, in order to come up with a memory array. The underlying aim here will be to inform the reader, future research scientist or development engineer, of some of the constraints affecting an elementary memory cell when it is to be integrated into an array.

### 15.2.1 General Principles of Matrix Storage

#### Structure of Information: Words and Bits

Whatever storage technology one may consider, the basic principles regarding the structure of information are always the same in existing memory products.

Data is stored in digital and binary form, and is structured in binary numbers (words) made up of a set of binary digits (bits equal to 0 or 1). An elementary cell in the memory plane corresponds to a single bit (0 or 1). Depending on the technology used, these two states are distinguished by two (metastable or stable) physical states such as:

- charge state of a capacitor (DRAM, EPROM, EEPROM),
- opening or closing of a switch (ROM),
- state of a logic gate (SRAM switch),
- orientation of polarisation (FeRAM) or magnetisation (MRAM),
- state of crystallinity (PCRAM),
- molecular conformation, spin orientation, etc.

In today's technology, and technology as it has been envisaged so far, these two possible states of a cell are detected by voltage or current levels, whatever the physical storage mode that is actually employed.

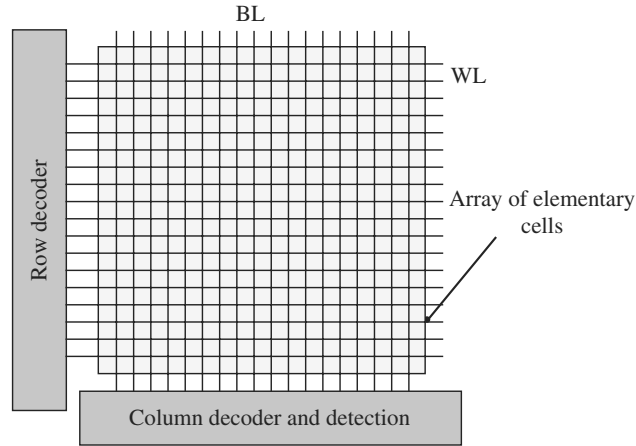
There are also multilevel memory cells. In this case, more than two logic states are stored in a single cell. It should be noted that this technique becomes exponentially complex, not only as far as storage is concerned, but also in terms of encoding and decoding. Indeed, to store 2 bits per memory cell, 4 logic levels must be discriminated; to store 3 bits, 8 levels must be discriminated; and so on. And the transition to 3 levels per cell, or some other non-multiple of 2, involves a significant complication in the decoding logic which is not cost-effective in terms of circuit area, at least so long as the rest of the system remains organised according to current architectures.

## Matrix Structure

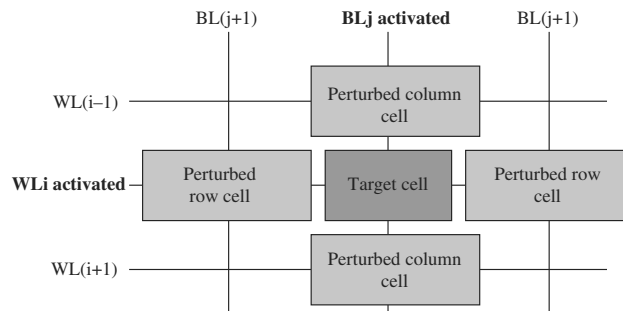
The memory plane contains a large number of cells arranged in the rows and columns of a matrix array on the surface of the circuit (the  $x, y$  plane in standard coordinates). A row of the array corresponds to a word or group of words (page) and a column to a bit. Each elementary cell is connected to two orthogonal conducting lines: the horizontal one is the word line (WL), common to all cells in the row, and the vertical one is the bit line (BL), shared by all cells in the column. Of course, these two lines are electrically independent. This means that they are constructed in different layers, i.e., at different depths  $z$ .

A single bit is accessed by simultaneously selecting a row and a column. Selection involves either imposing a specific voltage, or opening a switch to sample the current or voltage in the line (see Fig. 15.7).

Although the basic idea is very simple, this kind of organisation involves non-trivial properties of the memory cell. Indeed, a cell must react to the simultaneous selection of its row and its column connections, but remain insensitive to the selection of just one of these connections. This will become clearer when we detail read, write and erase strategies with reference to the perturbation of neighbouring cells. To achieve a matrix organisation, read,



**Fig. 15.7.** Underlying structure of a matrix memory. Row decoder, column decoder, and column detection system

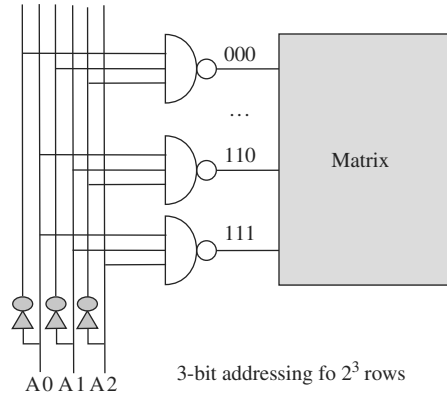


**Fig. 15.8.** Perturbation of neighbouring cells. When a cell is selected, all cells in the same row undergo a row perturbation and all cells in the same column, a column perturbation

write and erase processes requiring access to an elementary cell must satisfy the following criterion: access must be effective when BL and WL are simultaneously activated, but the perturbation when only one of these lines is activated must remain minimal. Otherwise, if this is not possible, all cells affected must be rewritten after reading (see Fig. 15.8).

Another accessing system that we shall only mention here for the record is used in charge transfer devices such as CCDs, or in shift registers. The idea here is to transfer data from one cell to the next, e.g., along a row (see Fig. 15.9). Data can only be accessed in whole lines, and reading or writing a line requires as many shifts as there are cells in the line. This approach is not therefore very fast and is no longer used in matrix memories.

The retention time of a cell is associated physically with the relaxation time of the state of the bit. The physical quantity stored becomes too degraded to be detected as it was recorded and the data is lost or inaccurate. Depending on



**Fig. 15.9.** Example of addressing: row decoder. Simplified logic diagram of a row decoder, with just 3 bits in this case. The column decoder is not very different in principle

the type of memory, the specifications for this time can vary from around 1 ms for refreshed circuits up to 10 or 20 years for so-called non-volatile memories. This time depends on the storage mode, but also on the read mode.

Furthermore, it should be stressed that, for a complete memory plane, the important feature, which determines either the reliability of the product or the required refresh frequency, is not the average relaxation time of a cell, but the relaxation time of the first failing bit cell in the whole plane! Therefore, in order to make high-capacity memory, it is not enough to make denser and/or bigger memory planes.

Note also that the operating temperature range within which circuit specifications can be guaranteed is another important constraint. For the general public, this range covers  $-40$  to  $+125^\circ\text{C}$ , extending up to  $150^\circ\text{C}$  for automotive applications. In more specialised contexts, the standards can be more severe, e.g., for military or space applications.

### Classification of Memories and Market Segmentation

Depending on their characteristics, memories can be grouped into three main families:

- Read only memory (ROM). Information is coded once and for all during fabrication of the circuit. This type of information can be used to start up a system when it is switched on.
- Random access memory (RAM). This is undoubtedly the best known and the most widespread form of memory. RAM specifications usually require very large sizes and fast write and erase access times. It must be possible to modify each bit at any time and independently of the others. Typically, this is where programs are stored when running, together with associated data.



- Non-volatile memory. The priority here is given to conserving data when the circuit is no longer switched on. These circuits are used to conserve personalised data between two work sessions (mobile phone directories, card codes, etc.).

In the past, memory chips corresponding to the distinct families were fabricated on different chips which were assembled on a card with a central processing unit that was itself on another chip containing only logic circuits. However, the trend towards miniaturisation and cost reduction has led to the idea of assembling as many distinct functions as possible on the same chip. Another consequence of this is a much faster communication time between functions, due to the shared bus. The trend has given rise to the idea of on-board memories, which come into their own in the market for small, highly differentiated and autonomous systems. Another decisive advantage of on-board memory is the reduced energy consumption of the system.

### Repair and Redundancy

For the reasons discussed above, the industrial fabrication of large matrices with suitable yields can be achieved today by integrating several clever pieces of design into the circuit itself, such as the possibility of replacing defective cells, or the insertion of redundant bits to correct a certain number of errors, i.e., physical repair or inclusion of logical redundancy:

- Repair of rows and columns found to be defective during fabrication. In a given array, a small number of extra rows and columns are included at the design stage. If some rows or columns turn out to be defective when tested off the production line, these can be replaced by rewiring or reprogramming. Rewiring is carried out by melting fuses, either with a laser or by electrical programming before delivering the chip.
- The use of error-correcting codes. To counter the problem of residual defective bits, write errors, or indeed untimely bit reversals, extra bits calculated on the data bits are added to words, so that a small number of defects can be corrected, or a possible error detected, at readout.

#### 15.2.2 Difficulties in Reducing Memory Cells to Nanoscale Sizes

In the following, we shall discuss the many technical difficulties associated with size reduction. However, before going into these details, it is worth spelling out the three main stumbling blocks.

#### Worsening Noise Effects

As the size is reduced, the amount of energy separating two states of an elementary cell also falls, making the system more and more sensitive to the

various sources of noise that might reverse the stored bit. Electromagnetic perturbations, interference<sup>2</sup>, or voltage variations across the terminals of the circuit all constitute noise sources that are difficult to eliminate. Array architectures have gradually been optimised to minimise the impact of noise (line crossovers, local comparisons, etc.), but signal reduction is not intrinsically favourable.

### Importance of Cell Connections with Outside Circuitry

Miniaturisation of the elementary data storage cell is not enough to miniaturise the full memory circuit, because it must be possible to address the cells. It is worth noting that, in the race towards miniaturisation, the limiting factor (and by a wide margin), even for pure logic circuits, is the routing of interconnects. The addressing network is thus crucial and the number of connections per cell becomes a basic criterion of integration density. In most cases, this number remains at 2 connections per cell, except in certain specific applications.

### Detection System

Once the electrical signal has been carried to the periphery of the circuit, it must be detected and correctly interpreted. Here again, the reduction in stored energy does not make things easier. Most architectures used today compare the detected signal with a local reference (differential detection), since fabrication processes induce a high level of variation in cell properties over a large array.

Note also that the testability and reliability of memory circuits are further constraints to be taken into account. Of course, the ideal (future) memory would be one combining all advantages simultaneously: very short access times (like SRAM), retention time in the non-volatile range, an aptitude for scale reduction in individual cells, possibility of increasing array size, immunity from external perturbation, etc. And if possible, all this for lower cost!

### 15.2.3 Matrix Memory Technology in Current Use

The general trends in the integrated memory market are not different from those of the global microelectronics market:

- low cost,
- low energy consumption,
- high densities and high performance.

<sup>2</sup> Perturbations can have various origins: cosmic rays, interference from other parts of the circuit, the environment of the system (in a car, aircraft, etc.), and so on.

Paradoxically, the rapid evolution of circuit performance makes it difficult to introduce radically new concepts. If a novel design is to make its way in this context, it must do better than existing designs, despite its lack of maturity. Moreover, a new concept must not only perform better than the  $N$ th generation, but it must also provide an easier transition to further generations. Finally, given the investment levels needed to develop any form of technology, levels that are in exponential growth, a break in the well-established flow proves very difficult and is only attempted when absolutely necessary. Nevertheless, the transition between new ideas and industrialisation has never been as fast as over the past few years in microelectronics.

### Non-Volatile Memory: ROM, EPROM, and Flash EEPROM

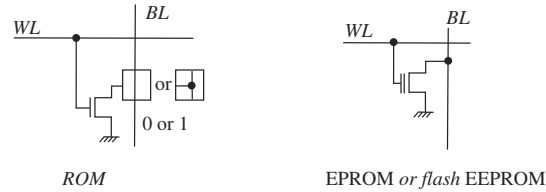
Read only memories (ROM) store information defined when the circuit itself is made, without the possibility of later modification. The individual cell is a single transistor and the 0/1 coding is simply associated with an on or an off state (see Fig. 15.10). This state can be defined by several distinct techniques:

- connection or otherwise of the transistor drain to the bit line (contact or no contact),
- insertion or otherwise of the transistor in the matrix,
- doping or otherwise of the channel.

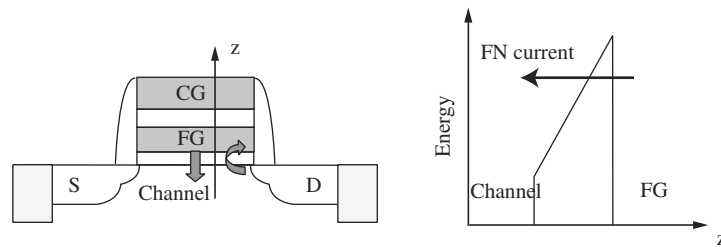
In order to be able to program and/or reprogram the matrix, one must use an electrically programmable read only memory (EPROM) or an electrically erasable programmable read only memory (EEPROM) (see Fig. 15.10) [9]. In the EPROM and the flash EEPROM, the elementary cell is a single transistor<sup>3</sup> comprising a control gate connected to the WL and a floating gate intercalated between the channel and the control gate. Writing in the floating gate is achieved by passing a strong current through this transistor. The high-energy carriers accelerated in this way induce ionisation (called impact ionisation) when they arrive at the sharp junction on the drain side of the transistor, whereupon the carriers get past the gate oxide of the transistor and charge up the floating gate. Depending on the charge state of the floating gate, this storage transistor will be on or off (states 0 and 1). The two states of the cell thus correspond to two different threshold voltages. Once the read voltages have been frozen, this amounts to an on or off state of the transistor.

In the EPROM, erasure is only possible by exposing the circuit to ultraviolet (UV) radiation for a long enough time (several minutes) to evacuate all charge stored in the floating gates of the array. Reprogramming thus requires an external physical intervention. The chip usually has to be disassembled for treatment in the UV oven, but the fabrication itself is not specific to the information stored and the circuit can be reused after UV erasure.

<sup>3</sup> Note that flash EEPROM, very similar to EPROM, differs from EEPROM which comprises electrically programmable cells but containing two transistors, one with a floating gate.



**Fig. 15.10.** Architectures of several non-volatile memory cells. A ROM cell is a transistor, connected or otherwise. An EPROM or a flash EEPROM comprises a single transistor with a floating gate integrated between the control gate and the channel. The EPROM is only reprogrammable after UV erasure, which requires specific packaging. To reduce costs, EPROMs have also been developed without this possibility of erasure. These are known as OTPs (one time programmable)



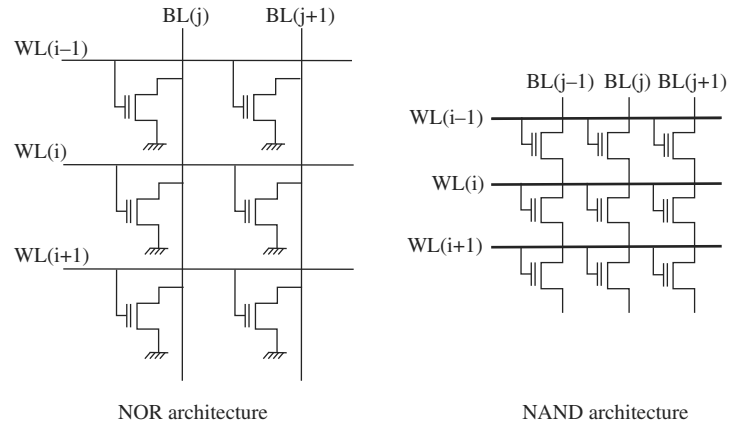
**Fig. 15.11.** Flash EEPROM cell. This comprises a transistor integrating a floating gate (FG) between the MOSFET channel and the control gate (CG). Writing is carried out by hot carriers from the drain. Erasure is achieved by field-assisted tunneling (Fowler–Nordheim effect): the field induces a reduction in the width of the energy barrier seen by the electrons, which can thus tunnel through it

In the flash EEPROM, this erasure is carried out electrically and in whole memory blocks. To do this, the control gate is placed under a high (negative) write voltage ( $< -10$  V), so as to deform the energy barrier represented by the dielectric. This is the Fowler–Nordheim effect (see Fig. 15.11). Writing is nevertheless rapid.

Erasure is collective but individual addressing in both write and erase is ensured by using several blocks. Hence, instead of modifying only a few cells in a block, the latter is recopied in a new block, taking into account the required modifications, before the old one is erased. This means that at least one block must remain free. For a memory of 16 blocks, the price to pay is thus 6% of the total capacity. Addressing is managed by an additional control software.

### Organisation of Non-Volatile Memory Planes: NAND vs. NOR

In the so-called NOR structure, each transistor has its gate connected to the WL and its drain to the BL (see Fig. 15.12). To improve densities, but to the general detriment of performance, it is better to use a NAND organisation in which all transistors of a given BL are connected in series. To read a cell A in a NAND



**Fig. 15.12.** Non-volatile NOR and NAND architectures. A NOR plane performs better but is less dense than a NAND plane, in which each cell does not have its own contact since the transistors are connected in series

architecture, all transistors in the BL are forced to conduct by a high voltage on the WL, except that of the cell A, to which an intermediate voltage is applied. The current through the series of transistors of the BL will thus be determined (1 or 0) solely by the cell A.

### Prospects for Non-Volatile Memories

This type of non-volatile memory is difficult to extrapolate for more than a few generations into the future<sup>4</sup> for the following reasons:

- The thickness of the gate dielectric which ensures retention cannot be reduced below 6 nm, otherwise it will be impossible to guarantee a retention time of 10 years, which is the standard. For reasons connected with the reliability of oxides, this limit is even fixed at 9 nm for polysilicon floating gates. This margin of reliability can be slightly enlarged by using materials that trap charges locally, preventing them from circulating within the floating gate. The latter can thus be made from silicon nitride [10], or as recently suggested, from semiconducting nanocrystals. Consequently, the retention of charge by a given memory cell is no longer governed purely by the weakest point of the dielectric, but results from the average retention in each floating nanogate, which is statistically favourable.

<sup>4</sup> In 2003, the 90-nm generations went into production. For the flash technologies, this corresponds to a gate length of 250 nm. The gate length has not been reduced below 0.25  $\mu\text{m}$  and has remained the same over several generations in the flash technologies. Only the rest of the design rules have been able to evolve, with subsequent increase in density and performance.

- Given that there is a floating gate and that the coupling between the floating gate and the control gate is inevitably imperfect, the applied gate voltage has a lesser effect compared with a logic transistor. One consequence of this is to greatly reduce control of the channel, giving rise to much more pronounced short channel effects (see Sect. 11.2.2). In 0.18- $\mu\text{m}$  technology (gate lengths of 0.25  $\mu\text{m}$  for flash memories), the required channel architectures are similar to those envisaged for much shorter logic transistors and at the feasibility limit for the approaches used today! As for the fabrication processes of logic transistors, it is hoped that high permittivity dielectrics will be able to prolong the lifetime of these concepts to some extent. Having said this, it is difficult to see clearly beyond the next generation of this technology, not for technical reasons, but for much more fundamental physical reasons.

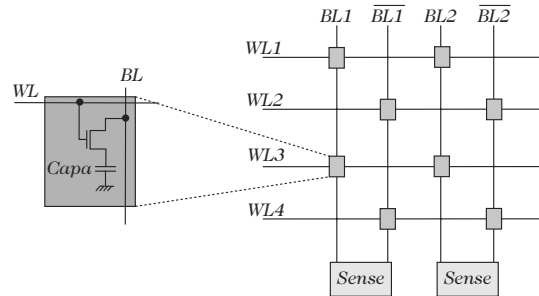
### Volatile Memories: RAM, DRAM, and SRAM

RAM memories are volatile. In contrast to non-volatile memory, information is only conserved whilst the circuit is switched on. On the other hand, access times are much shorter.

A dynamic random access memory (DRAM) cell comprises an access transistor and a charge storage capacitor. Unlike the ROM, EPROM or EEPROM memories, the energy associated with the detected signal is not supplied by the peripherals at the time of reading, but is directly paid back by the discharge of the capacitor when the transistor is switched on. (Charge is transferred from the previously charged capacitor in the BL, which induces a differentially amplified change in the voltage.) Characteristic charging and discharging times are very short: 10–20 ns for a purely random access cycle, and a few ns/bit when a whole page is accessed. A refresh procedure is required to ensure the survival of the information for retention times of 10–100 ms. Reading is destructive and requires rewriting.<sup>5</sup>

Storage capacity is less than the capacity of the BL itself, in the ratio of 1 to 7. When the access transistor is switched on (by applying a voltage across the WL), stored charge must be detected by comparison with a reference BL. This detection is achieved by differential amplifiers connected at the end of the BL and comparing the read BL with a reference BL. In folded BL structures (see Fig. 15.13), the reference BL is nothing other than the neighbouring BL (complementary BL) in which a specific cell containing a reference value (charged capacitor) is activated. This organisation protects against local voltage variations in the memory plane. In certain cases, the memory structure includes a cache made from SRAM cells which allow one to store a whole page and modify this page if necessary before rewriting it in the array.

<sup>5</sup> Reading is destructive because the charge contained in the capacitor is shared in the BL, but this charge is automatically restored (rewritten) during the read process. Hence, the program does not necessarily have to implement a temporary memorisation of the data read following a true rewrite.



**Fig. 15.13.** Organisation of a DRAM memory plane with folded bit line. The elementary cell is made from a transistor by connecting a capacitor to the BL. The differential amplifier compares two neighbouring BL, only one of which contains an activated cell

Static random access memories (SRAM) are made from transistors using the same fabrication process as for logic gates, and thus involve no extra fabrication stage compared with the fabrication of logic circuits. Another advantage is that they have faster access times than DRAM.<sup>6</sup> However, this solution is much less dense (6 transistors per cell) and implies a higher power consumption.<sup>7</sup> The logical principle is as follows: two inverters (each made from an N transistor and a P transistor) are connected head to tail, to make a logic switch connected by two access transistors to the BL and complementary BL. The state of the cell is read by comparing the BL and complementary BL by means of the differential amplifier. This arrangement requires two BL per cell, which is highly efficient for compensating offsets, but at the same time takes up a lot of space. This extra BL can be suppressed by comparing the BL with a reference rather than the complementary BL.

### Prospects for RAM

In the last few DRAM generations, size has been reduced without significant change in the storage capacity (30 fF). This has been achieved by integrating the latter in 3D with ever higher aspect ratios (50 in 2003, for a width of 120 nm and a height of 6–8  $\mu\text{m}$ ). However, such a non-homothetic reduction of the various dimensions soon reaches its limits. To compensate in part for the reduction of space,<sup>8</sup> high permittivity dielectrics are used, including oxynitrides, nitrides,  $\text{Ta}_2\text{O}_5$ ,  $\text{SrTiO}_3$ ,  $\text{Al}_2\text{O}_3$ , and  $\text{HfO}_2$ , to increase the areal capacitance of the dielectric without altering the leakage from it, which leads

<sup>6</sup> In random access, the typical cycle time of an SRAM is a few nanoseconds.

<sup>7</sup> A distinction is made between static consumption (without access) and dynamic consumption (with access). The dynamic consumption of DRAM is typically only half that of SRAM.

<sup>8</sup> Use of a very rough electrode material can extend the developed surface of the capacitor at fixed area.

to a loss of retention of the cell. Little by little, design efforts have reduced the capacitance detectable by the amplifiers to an absolute minimum.

The problem when reducing the size of SRAM is directly related to the reduction in size of the logic gates, which in itself may be seen rather as an advantage over the EEPROM or the DRAM. However, sensitivity to noise is a very important problem.

In the context of the future ideal memory, viz., a non-volatile and fast RAM, it is worth mentioning that the standards have been set: the NOVO-RAM memory combines the structure of an SRAM with the non-volatility of an EEPROM. When active, information is stored in the bistable SRAM. When the system is switched off, the stored data is transferred locally into an EEPROM structure located directly in the cell. The advantage is a very fast transfer, because volatile and non-volatile storage are completely parallel, since carried out simultaneously for all cells in the memory array. However, the elementary cell remains bulky.

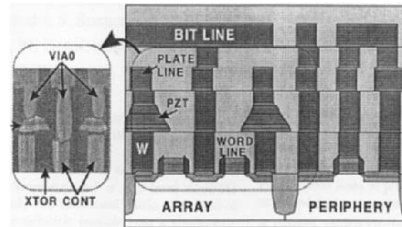
#### 15.2.4 Memory Concepts Under Development

FeRAM, MRAM and PCRAM are all justly presented as combining the advantages of non-volatility (as for flash EEPROM), random access and very short access times (as for SRAM), and high integration densities (as for DRAM). A further point that remains to be demonstrated for these concepts concerns the ease with which they can be integrated into current fabrication processes for logic circuits. This is why there is still room for alternative ideas, perhaps less ideal in terms of performance, but easily integrated into current or future fabrication technologies for logic circuits.

#### FeRAM Memories

The underlying principle of the FeRAM memories, as for DRAM, combines a capacitor and an access transistor (see Fig. 15.14). However, in the case of the FeRAM, the dielectric of the capacitor exhibits a bistable state of electrical polarisation which allows retention times in the non-volatile range. These dielectric materials, such as PZT ( $\text{Pb}_x\text{Zr}_{1-x}\text{TiO}_3$ ) or SBT ( $\text{SrBi}_2\text{Ta}_2\text{O}_9$ ) display ferroelectric behaviour with distinct hysteresis in the polarisation–voltage curve. PZT was the first studied and has the advantage of a better polarisation remanence, as well as lower fabrication temperatures. On the other hand, SBT has the advantage of behaving better over long periods and a better aptitude for size reduction. In a FeRAM cell, writing is carried out by applying a voltage across the capacitor. Reading remains destructive since it consists in detecting the displacement current associated with the polarisation reversal. A key point in the study of ferroelectric dielectric materials is the gradual loss of their properties after repeated write/erase cycles. Indeed, the hysteresis cycle changes shape, either shrinking or distorting.





**Fig. 15.14.** FeRAM memory. The capacitor is connected via an access transistor to the bit line, as in a DRAM. (Presented at the 2003 VLSI conference by Texas Instruments)

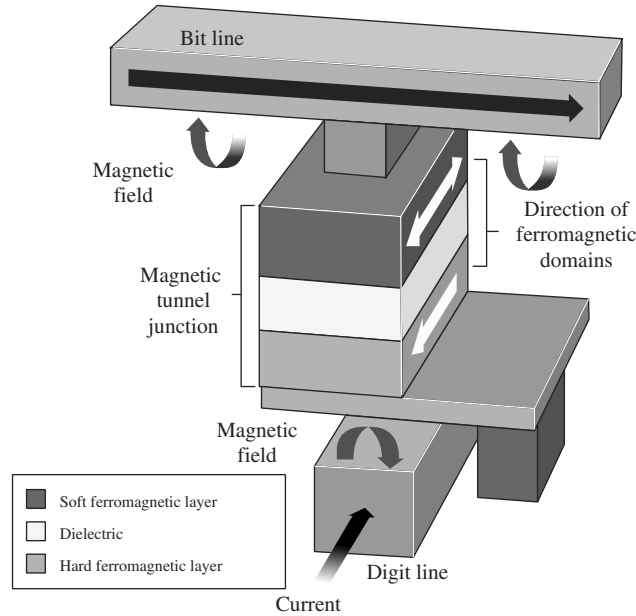
This kind of memory has proven its ability for very fast write times ( $< 100$  ns) with very small voltages and power inputs. The main advantage is, however, the very low operating voltage. The main difficulty with FeRAM design resides in the dielectric material itself and its integration into standard microelectronic fabrication processes. For the time being, demonstrations have not been competitive in terms of required space<sup>9</sup> and prospects for size reduction with this design do not look good. Another weak point is the endurance of the material. However, there is a market for certain highly specific applications, such as automobile black boxes.

### MRAM and PCRAM Memories

In the way they are organised, these concepts bear a certain resemblance, since in both cases, information is detected by the change in resistivity in an elementary layer (EL) located at the intersection of two metal lines. To detect this change in resistivity, a diode, or preferably a transistor, must be added in series with the variable resistance. Indeed, indirect conduction channels passing through several neighbouring cells must be avoided, because they would vitiate the measurement of the resistivity in the cell to be read. This transistor is connected to the BL via the EL and also to the source line.

From a physical point of view, however, the information is stored very differently in the two cases. In the magnetic random access memory (MRAM) the change in resistivity is a consequence of a change in the magnetic orientation between neighbouring ferromagnetic layers (see Fig. 15.15). The resistive layer is a magnetic tunnel junction (MTJ). The change in resistivity of an MTJ is a consequence of the spin-dependent tunnel effect. In its simplest version, the MTJ comprises a dielectric tunnel sandwiched between two ferromagnetic layers. The upper layer has an easily reversible magnetisation and

<sup>9</sup> The ITRS Roadmap indicates an elementary cell of  $5\mu\text{m}^2$  for FeRAM in 2002. However, technology is evolving very fast and Texas Instruments recently published (IEDM 2003) a prototype at  $0.53\mu\text{m}^2/\text{cell}$  (64 Mbit plane).



**Fig. 15.15.** Schematic of an MRAM memory cell. The stack of magnetic materials is placed between the BL and an underlying access transistor for reading the data. To write data, it is the BL and the digit line that are activated

is a permalloy such as NiFe/CoFe. The lower layer has a fixed or pinned magnetisation and is made from a material with high coercive field, such as CoFe. In practice, an antiferromagnetic layer of FeMn, IrMn, or CrPtMn with zero overall magnetisation is included beneath the lower (soft) layer to provide a reference direction for the hard layer.

In this cell, writing consists in flipping the spin of the soft layer by creating a local magnetic field, a consequence of the current imposed in the metal lines. For selective writing in a given cell, the high and low lines are activated simultaneously with a current chosen in such a way that the field  $H_x$  generated by each line obeys

$$H_x\sqrt{2} > H_c > H_x, \quad \text{where } H_c \text{ is the reversal field.}$$

In this way, only the target cell is flipped and neighbouring cells retain the same orientation.<sup>10</sup>

In the phase-change random access memory (PCRAM), the change in resistivity results from a phase change in a material that is conducting in its crystalline phase and insulating in its amorphous phase. These phase-change

<sup>10</sup> This condition is hardly fulfilled in real circuits with a high degree of reliability. This is one of the reasons why, in real circuitry, the write strategy will be more complex. However, the complete description of these strategies goes beyond the scope of this book.

materials are alloys based on group VI elements of the periodic table, known as chalcogenides, e.g.,  $\text{Ge}_2\text{Sb}_2\text{Te}_5$ . Writing into either the crystallised or the amorphous state is induced by the Joule effect, caused by passing a large current through the material. When writing, the current must be adjusted to ensure accurate control of the temperature and the local anneal time of the material. Transition to the crystallised state (state 1 by convention) is slower than the opposite.

Let us briefly compare the advantages and disadvantages of these two memories. Apart from their non-volatility, one advantage common to both MRAM and PCRAM is the non-destructive read mode.

In MRAM, writing is fast and can in principle<sup>11</sup> be done at low voltage. Endurance is also good. On the other hand, the fabrication process for the magnetic layers is delicate, and it remains to be shown that these devices are susceptible to size reduction. Write currents are large (with significant transients) and the detected signal is weak. Integration involves more than two connecting lines per bit (indeed, three lines are required, together with a local interconnect), which thwarts the move to higher integration densities.

In PCRAM, writing is also fast (10–100 ns), low voltage (3 V) and low power. Endurance is very good, random access guaranteed, and the size of the memory cell is small and well suited to scale reduction. The signal discrimination margin is also good. Note that these materials are used for a similar property in today's DVDs, for the purposes of optical refraction. They are therefore well known in the industrial context, although not yet integrated into silicon-based microelectronics. As for FeRAM and MRAM, the integration of a new material raises the usual difficulties: acceptability on the production line (contamination), compatibility with process temperatures, chemical compatibility, filling capacity for deposition on non-planar layers, availability of etch and/or chemomechanical polishing (CMP) processes, and so on. Finally, although highly promising, this concept also suffers from a certain difficulty in detecting the signal. We shall see below that other concepts, often somewhat futuristic, attempt to include a local amplification of the signal within the memory cell itself.

### **An Evolutionary Concept: 1T-Capacitorless DRAM**

The idea here is to use a phenomenon that has until now been considered as an undesired effect which manifests itself in SOI transistors and goes by the name of the kink effect. When a silicon-on-insulator (SOI) transistor is operating, charges generated by impact ionisation on the drain side of the sharp junction, which cannot escape via the lower part of the wafer as they

<sup>11</sup> This holds on the level of an individual cell, but in practice, since large write currents are required and lines do not have zero resistance, it is difficult to reduce the voltages in the periphery of the array.

might on a non-SOI wafer, accumulate under the transistor channel. As a consequence, the threshold voltage of the transistor will drift.<sup>12</sup>

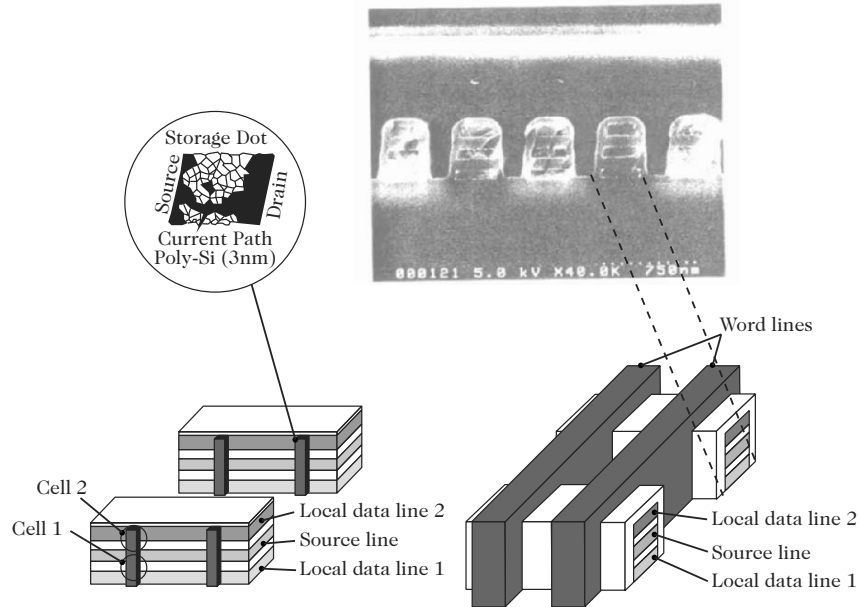
According to this proposal [11], the idea is to code data precisely by storing charge in the body of an SOI transistor. As in flash EEPROM, reading would consist in detecting the  $V_t$  of the transistor. Writing would be carried out by accumulating charge: a large source–drain current would generate impact ionisation on the drain side of the junction. Erasure, i.e., writing 0, would be obtained by applying a negative S/D voltage which would induce a bulk current in the body. Characteristic relaxation times for stored charges place this type of memory cell in the volatile category. Compared with a typical DRAM, one may bypass the need for any storage capacity, thereby removing the most difficult point. However, it will only become possible to experiment with this concept on a realistic scale when SOI technologies have been adopted industrially.

### Advanced Charge Storage Memories

A very good example of an advanced memory concept has been proposed by K. Yano (see Fig. 15.16). The unit cell comprises a transistor in which the channel is an ultrathin polysilicon film, with thickness of the order of 2–3 nm. The electrical properties of this film are dictated by a large number of local variations in the potential, consequences of the irregularity of the material. One can give a rather pictorial explanation of the physics of conduction and trapping. Electrons crossing this transistor encounter a mountainous potential landscape on the microscopic scale, including peaks and valleys. The idea is that, although this relief may well be different in each cell, there is always a low energy conduction path. In the vicinity of this conduction valley, there are trapping zones (plateaus and lakes) in which charges can become trapped when the Fermi level climbs then falls back. (The ‘rising waters’ correspond here to the application of a high gate/source voltage.) Once trapped, these charges will influence conduction in the valley, i.e., they will play the role of floating gates, modifying the threshold voltage of the conduction channel. In a certain sense, this cell is a nanoscale replica of an EEPROM cell.

As the channel is deposited, a vertical integration of the transistor is possible. K. Yano was able to fabricate a 128-kb memory plane by stacking alternate layers of conductors and insulators. This stack was then etched to define the conducting lines, one above the other, which define the sources and drains of vertical transistors. The polysilicon channel, the gate oxide, and the gate were then deposited and lithographed to define the WL, oriented in the direction perpendicular to the S/D lines. With three metal lines stacked vertically, two

<sup>12</sup> This is the equivalent of a substrate effect for a conventional (non-SOI) MOS transistor: the polarisation of the SOI (body) substrate is not controlled since it is floating. The change in potential of the body resulting from charge accumulation modifies the  $V_t$  of the transistor.



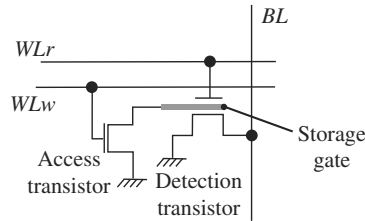
**Fig. 15.16.** Single-electron memory due to K. Yano. Storage and detection are carried out in an ultrathin polysilicon layer. The vertical integration of several S/D conducting lines can greatly increase the density [13]

double transistors are thus defined at each crossover of the WL and the S/D lines. In principle, one might contemplate increasing the number of cells per crossover: hence, for four superposed lines, one would obtain three cells, and for  $N$  lines,  $N - 1$  cells, which would achieve a very high integration density.

In practice, the main difficulty with this concept relates to the elementary cell itself. Indeed, write, read and erase voltage conditions are highly variable from one cell to the next, making addressing a delicate matter over the whole array. However, the high write speed is obtained here by the fact that charge storage occurs through a potential barrier that is not crossed, but which is rather overcome by modifying the Fermi level. No doubt an idea worth remembering!

### Various Versions of Gain Cells

The idea behind the gain cell is to combine the flexibility of an access transistor (as in DRAM) with amplified charge detection via a detection transistor (as in EEPROM) (see Fig. 15.17). Note that this is not a new idea, but has spurred current DRAM evolution. Having said this, one or other of these transistors can be either a field-effect transistor (FET) or, in a more futuristic context, a single-electron transistor (SET). It is the access transistor that is the most



**Fig. 15.17.** Two-transistor gain cell. The access transistor controls retention and writing in the storage gate and the latter controls the detection transistor. This cell architecture has the disadvantage of requiring two WLS, one to control reading ( $WL_r$ ) and one to control access ( $WL_w$ )

critical, because some way must be found to reduce its leakage current and maintain sufficient charge retention in the floating gate.

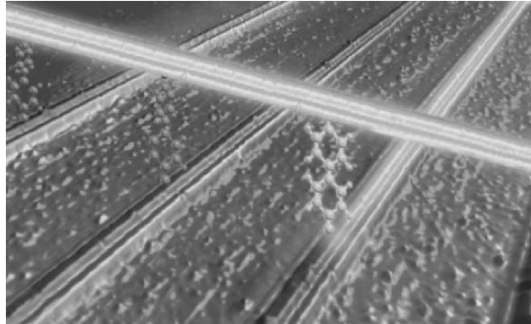
An example of gain cells combining two FET has been proposed at the University of Cambridge (UK). This is the so-called phase-state low electron number drive (PLED) memory. The access transistor is integrated vertically above the detection transistor. However, this vertical transistor is modified by inserting tunnel junctions and finely tuning the doping of the gate and channel so as to reduce the current in the on state in favour of a significantly reduced leakage current in the off state.

A second example has also been proposed by K. Yano. The access transistor is made on a very thin SOI layer (several nanometers), once again in such a way as to reduce the off current. One consequence of this reduction in the channel thickness is to increase the band gap, which reduces leakage by recombination generated in the channel. This transistor is known as the SESO (single-electron shut-off).

A third example, still more futuristic, has been put forward by A. Ahmed and coworkers at the University of Cambridge (UK). The access transistor is a SET, once again made in a thin SOI layer. The detection transistor is a FET comprising several gates in series. One is a charge storage gate connected to the access transistor and the other is a control gate. The main problem of the SET is its operating temperature, which remains in the cryogenic range ( $< 30$  K). When it becomes possible to make a room temperature SET, this idea will be worth taking up again.

Such a SET might perhaps be obtained by means of a vertical stack of tunnel dielectric layers and layers of nanocrystals. One proposal for a compact memory cell has been made along these lines by CEA-LETI in France. This cell architecture is inspired by flash memories, but writing in the floating gate is controlled by multiple tunnel junctions. In this situation, one exploits the non-linearity of the junction due to Coulomb blockade. It remains to demonstrate these effects at room temperature.

Still more futuristic ideas have been put forward, combining an access transistor and a detection transistor, both of which are SET [12]. Laboratory



**Fig. 15.18.** Matrix memory of the future. Carbon nanotubes crossing one another at right angles connect bistable molecules. (Artist's view, Hewlett Packard)

investigations have only reached the elementary cell level and operating temperatures remain very low indeed ( $< 10\text{ K}$ ). This cell does not really deserve the title of gain cell, for one of the weaknesses of the SET is precisely its low gain. However, this construction certainly prefigures ideas that will eventually see the light of day.

In all cases, the main stumbling block remains the access transistor, whose performance, in particular the ratio of the 'on' current to the 'off' current ( $I_{\text{on}}/I_{\text{off}}$ ), needs a significant boost. The search is on for a more efficient switch!

### Atomic Scale Elementary Cells

From the above discussion, the reader will no doubt have become aware of the constraints involved in integrating a very high capacity memory. One of the main points to remember is the limiting factor imposed by the problem of addressing individual cells. Some have already envisaged the possibility of reducing the BL and WL down to the atomic scale. At Hewlett Packard, a recent report suggests integrating a cell with two conduction states made from a bistable molecule and connected to two crossed carbon nanotubes. When we are able to exercise perfect control over nanotube growth, it may be possible to envisage orthogonal line structures (just like the arrangement currently proposed for MRAM and PCRAM) in which the variable resistance will be reduced to the molecular scale. But decoding and detection systems must then also be brought down to this scale, otherwise integration will clearly be impossible. To reduce a memory array, every one of its constituent elements must be reduced: cells, connecting lines, signal decoders and amplifiers, and also the greater part of the peripheral circuits. So lead on, molecular electronics!

### 15.3 Conclusion

Given the continued advance in performance and density, in both mass memory and matrix memory, a question arises: will these two types of memory, so different at the present time, ever converge? At the moment, they serve quite distinct market requirements with respect to available memory capacity and cost per Mbit.

In the short term, the hard disk is expected to evolve by implementing hybrid techniques (thermally assisted recording) and using perpendicularly magnetised storage media. Beyond this, over the next 5 to 10 years, the future begins to look more of an adventure. Will there be a quick transition to discrete media, or will microprobe techniques move forward fast enough to impose themselves? Matrix memories are also in rapid evolution. In a few years from now, volatile memories like DRAM, SRAM, and so on, will very probably be replaced by memories that are just as rapid but non-volatile, like FRAM, MRAM, or PCRAM.

In the longer term, some analysts raise the question of convergence between these two main types of memory. Indeed, mass memories (hard disk) and matrix memories are distinguished in three main aspects:

- density, two to three orders of magnitude higher in hard disks,
- time required to access data and data flow rate (read, write, etc.), which favour matrix memories in the case of random access,
- cost per Mbit, currently much lower for hard disks.

The place occupied by the various technologies in the world of tomorrow depends mainly on the way their performance evolves with regard to these three points. It is no easy matter to make accurate predictions when the technological breakthroughs required to step into the nanoworld have not yet been made. Nonetheless, certain trends, determined by market needs, are already clearly apparent: the matrix memories that will eventually dominate will be based on non-volatile technology, and microprobe memories will only come into their own if the cost per Mbit remains much lower than that of the matrix memories, which implies very high densities.

Other more structural developments are likely. Some functions currently handled by a mass memory will be entrusted to matrix memory. This prospect fits in with the general increase in the share of solid memories. It is supported by the market trend towards more diffuse microelectronics, i.e., towards small, more autonomous systems called systems on-chip (SoC), in which the integration of mechanical elements is significantly more problematic.

If we look still further into the future, we must emphasise that more radical developments are nevertheless possible. Hence, in today's architectures, logic functions remain well separated from memory functions, which is not at all what happens in biological systems such as the human brain. It is reasonable to think that the very notion of memory storage as we understand it today might evolve towards systems in which memory and logic are more intimately mixed



together. The addressing problem could then be solved, because information would be stored directly in the place it was needed. However, such an idea is still in the realm of science fiction!

## Appendix

The following is a brief guide to some of the technical terms used in this field.

*Access Time.* For a memory circuit, this is the time lag between a request for a stored bit and the moment it becomes available to external circuits. One way to reduce the access time, is to reduce array sizes, which works against the push towards high capacity matrices. This conflict leads to the arrangement of many subarrays within a large circuit. Delicate optimisation is then required at the design stage, usually highly specific to the application.

*Coercive Field.* In a hysteresis cycle  $M(H)$ , this is the field  $H$  corresponding to zero magnetisation  $M$ . For a square cycle, the ideal shape for magnetic storage media, it is the field that must be applied to reverse the magnetisation direction.

*Demagnetising Field.* A magnetic object produces a field within itself. Simple diagrams show that this field is always opposed to the magnetisation of the object, which it thus tends to destabilise, whence the term used.

*Design.* In microelectronics, this refers to the whole process of reflection and realisation of a detailed circuit diagram. The notion covers a wide range of tasks from the conception of systems bringing together macros which are themselves made up of more elementary functions, right down to drawing up plans for the elements to be fabricated. Over the last few years, there has been a spectacular level of specialisation in the job of the designer. But paradoxically, circuit optimisation today is the fruit of a detailed dialogue between designers and engineers in charge of the physical aspects of making silicon circuits.

*Field Effect Transistor (FET).* This is the most commonly used transistor. Its channel connects a source to a drain via a (semiconducting) channel controlled by a gate. Applying a voltage to the gate modifies the conduction state of the channel electrostatically and the transistor goes from the off state to the on state.

*Giant Magnetoresistance.* Variation of electrical resistance observed in a trilayer composed of two magnetic layers separated by a non-magnetic thickness depending on the relative magnetisation directions of the two magnetic layers (see Chap. 14).

*Integration Density.* One of the main goals in microelectronics is to reduce costs. The most efficient way to do this is to reduce circuit sizes in order to fabricate more on a slice of silicon. Increasing the integration density means increasing the number of functions integrated on a given area, i.e., designing circuits with smaller elementary patterns and also optimising the way circuits are arranged. In parallel with this, fabricated circuits are more and more complex and involve ever more elementary transistors. In the end, despite continued growth in the diameters of silicon wafers, there are fewer chips per wafer, at least where the most highly evolved circuits are concerned.

*L1<sub>0</sub> Alloys.* These alloys, e.g., FePt, FePd, are currently attracting a great deal of interest due to their remarkably high magnetocrystalline anisotropy, which arises from the chemical ordering in the so-called tetragonal L1<sub>0</sub> structure. (The chemically disordered phase does not exhibit this anisotropy.) Such a high level of anisotropy can be used to stabilise the magnetisation of very small ferromagnetic grains, whence the possible use of these alloys in future magnetic storage media.

*Magnetic Anisotropy.* The anisotropy of a magnetic particle reflects the existence of some favoured direction for magnetisation. It arises either from the shape of the particle, or from the properties of the material. In the first case, we speak of shape anisotropy. An example is provided by the spontaneous orientation of the magnetisation in a needle along its principal axis. In the second case, we speak of magnetocrystalline anisotropy, since it is related to the presence of a crystal lattice. In today's hard disks, adding platinum to the alloy brings about this kind of anisotropy by deforming the initially cubic lattice of cobalt in a tetragonal manner.

*Magnetoresistive Read Heads.* These use the phenomenon of anisotropic magnetoresistance: the electrical resistivity of a ferromagnetic material depends on the angle between the magnetisation and the current, e.g., for a permalloy, the resistivity is about 2% higher at room temperature for a magnetisation parallel to the current. A magnetoresistive read head comprises a soft magnetic layer with a current through it, whose magnetisation direction and hence whose electrical resistance is modified under the effects of the leakage field from magnetic transitions. Since 1999, these heads have been replaced in hard disks by giant magnetoresistive read heads.

*Recording Density.* This expresses the quantity of data stored per unit area of a 2D medium, in bit/cm<sup>2</sup>, where 1 bit is one bit of elementary data, i.e., 0 or 1. The international literature often uses the inch (1 inch = 2.54 cm). It is important not to confuse the bit and the byte (1 byte = 8 bit). In 2003, an up-to-date hard disk has a capacity of the order of 800 Gbit and a density of the order of 20 Gbit/in<sup>2</sup>.

*Retention Time.* This is the maximum time during which information stored in an elementary cell remains unblemished. It depends on the intrinsic performance of the cell, but also the discrimination threshold between 0 and 1.

Note that, for a matrix memory, the retention time is not the average retention time taken over all cells, but rather the retention time of the weakest cell. The retention time of a matrix thus depends on its size.

*RKKY Coupling.* In a normal, non-ferromagnetic metal, the electron population is not polarised, i.e., as many electrons carry spin up as spin down. However, in contact with a ferromagnetic metal, the two populations become unbalanced by a proximity effect. This imbalance gradually falls off with the distance from the ferromagnetic interface, by oscillation in the spin populations. RKKY coupling, named after its discoverers Ruderman, Kittel, Kasuya and Yosida, results from this oscillation. If the non-magnetic layer separating two ferromagnetic materials has just the thickness to ensure that the spin populations on its two interfaces are opposite, it induces a spontaneous anti-ferromagnetic coupling.

*Shape Anisotropy.* The magnetic field which a magnetic object creates within itself is in the opposite direction to its own magnetisation. To minimise this unfavourable interaction, and in the absence of other sources of anisotropy, the magnetisation of an ellipsoidal particle orients itself spontaneously along its major axis, whilst the magnetisation of a thin magnetic film orients itself spontaneously in the plane of the thin layer.

*Single Electron Transistor (SET).* This also operates under electrostatic control but exploits carrier charge quantisation. The conduction channel comprises at least one island separated from the source and drain by tunnel junctions. The electrostatic potential of the island is controlled by a gate. Although it remains a research device with no proven industrial applications, this may change, especially if it can be made to work at room temperature.

*Soft Material.* Magnetic material with weak coercive field, i.e., its magnetisation direction reverses under application of very weak magnetic fields ( $2\text{--}3 \times 10^{-4}$  T for the permalloy  $\text{Fe}_{80}\text{Ni}_{20}$ ).

*Superparamagnetism.* On the atomic scale, a ferromagnetic material is characterised by parallel alignment of spins at neighbouring sites of the crystal lattice, which gives rise to a characteristic bulk magnetisation  $M_s$  in the material. Above the Curie temperature, thermal fluctuations dominate over exchange between neighbouring spins and the spin directions begin to fluctuate randomly, thereby cancelling the average magnetisation (paramagnetic state). Superparamagnetism corresponds to an identical phenomenon on the scale of a particle with volume  $V$  and anisotropy  $K_u$ . Indeed, the energy barrier between the two magnetisation directions has height  $K_u V$ . For small volumes,  $K_u V \sim k_B T$  and thermal fluctuations are large enough to carry the particle magnetisation along with them. It is important to note that the particle nevertheless remains ferromagnetic, i.e., its overall magnetisation fluctuates, but the ferromagnetic order of the spins is preserved.

*Write Time.* This is the time required to record the information for storage. Like the access time, it is highly variable, depending on whether the relevant bits are isolated in the matrix or well located in a contiguous manner on the same row. One speaks of random access in the first case and page or burst mode in the second.

## References

1. Zhang, Z., et al.: IEEE Trans. Magn. **38**, No. 5, 1861 (2002)
2. Weller, D., and Moser, A.: IEEE Trans. Magn. **35**, 4423 (1999)
3. Fullerton, E.E., Margulies, D.T., Supper, N., Do, H., Schabes M., Berger, A., and Moser, A.: IEEE Trans. Magn. **39**, No. 2, 639 (2003)
4. Wood, R., Sonobe, Y., Jin Z., and Wilson, B.: J. Magn. Magn. Mat. **235**, 1 (2001)
5. Wu, W., Bo, C., Xiao, Y.S., Wei, Z., Lei, Z., Linshu, K., Chou, S.U.: J. Vac. Sci. Tech. B **16**, No. 6, 3825 (1998)
6. Sun, S., Murray, C.B., Weller, D., Folks, L., Moser, A.: Science **287**, 1989 (2000)
7. Chappert, C., Bernas, H., Ferré, J., Kottler, V., Jamet, J.P., Chen, Y., Cambriil, E., Devolder, T., Rousseaux, F., Mathet, V., Launois, H.: Science **280**, 1919 (1998)
8. Vettiger, P., Cross, G., Despont, M., Drechsler, U., Dürig, U., Gotsmann, B., Häberle, W., Lantz, M.A., Rothuizen, H.E., Stutz, R., Binnig, G.K.: IEEE Trans. Nanotech. **1**, No. 1, 39 (2002); Vettiger, P., Despont, M., Drechsler, U., Dürig, U., Häberle, W., Lutwyche, M.I., Rothuizen, H.E., Stutz, R.: IBM J. Res. Dev. **44**, 323 (2000)
9. Cappelletti, P., Golla, C., Olivo, P., Zanoni, E.: *Flash Memories*, Kluwer Academic Publishers (Boston, Dordrecht, London, 1999)
10. Eitan, B., et al.: IEEE Elect. Dev. Lett. **21**, 11, 543 (2000)
11. Fazan, P., et al.: IEEE Int. SOI Conference 2002. See also: Ohsawa, T., et al. (Toshiba): IEEE J. Solid-State Circuits **37**, 11, 1510 (2002)
12. Dresselhaus, P.D., et al.: Phys. Rev. Lett. **72**, 20, 3326 (1994). See also: Lafarge et al.: CRAC **314**, 883 (1992)
13. Yano, K.: IEE **87**, No. 4 (April 1999)

## Optronics

J.-L. Pautrat, J.-M. Gérard, É. Bustarret, D. Cassagne, E. Hadji, and C. Seassal

Optronics is a contraction of optics and electronics. Its aim is to exploit technological developments across the board to bring these two sciences together. Techniques for structuring materials, whether metallic, insulating, or semiconductor, have led to new fields of activity, because the properties of materials can be modified and new materials tailor-made for specific applications. There are just too many different phenomena involved and too many fields of application to tackle them all in detail here. We shall therefore aim only to present the most advanced, such as optical sieves, semiconductor quantum dots, and photonic band gap (PBG) materials.

### 16.1 Surface Plasmons and Nanoscale Optics

#### 16.1.1 Introduction

The most fundamental limitation facing technological efforts to miniaturise optical sources, detectors or data-processing devices is the Rayleigh criterion. This says that efficient propagation of radiative electromagnetic modes, as achieved in fibre optics, is limited by diffraction phenomena to wave guides or apertures with transverse dimensions greater than half the effective wavelength in the relevant medium, i.e., several tenths of a micron in the case of visible light. This is also the ultimate dimension for focussing a light beam in a classical microscope (Abbe criterion).

Another approach to the miniaturisation of optical functions is inspired by the way high frequency electrical signals can be guided in metallic lines with thicknesses much less than the wavelength. The main drawback is that the electrical conductivity of metals is much lower at light frequencies (1 000 THz) than at 100 MHz, for example. This means that signals can only be propagated over a few microns. However, in this context, there are other phenomena at light frequencies which allow one to overcome the difficulty.

One particularly promising idea is to use coherent oscillations of electrical charges on the surface of metals or ionic crystals. Indeed, the electromagnetic field associated with these longitudinal waves is strongly confined at the metal/air interface and they have a high degree of spatial coherence. This means that circuits and planar devices can be devised on length scales well below the wavelength. We shall see that techniques for structuring these materials on a scale of around 10 nm are being developed in parallel with means for local observation and numerical modelling of these non-radiative waves.

Collective oscillations of surface charges are called surface plasmons, or plasmon polaritons when one wishes to emphasise their coupling with an external electromagnetic field. There are two types of plasmon:

- plasmons guided by the surface of thin, flat metallic layers, propagating with a phase velocity of the same order as the speed of light,
- plasmons found in metal particles of subwavelength dimensions, which are in fact dipole electron oscillations bounded by the nanoscopic particle.

In the following, we begin by defining surface plasmons more precisely and stating the main laws governing their behaviour. We shall emphasise the coupling of such modes with an external electromagnetic field and their great sensitivity to local surface features. We shall then describe a certain number of applications exploiting these properties, especially in the fields of spectroscopy and detection, and which appeal increasingly to techniques of integrated optics on the micron scale. We then turn to the case of metal surfaces with high aspect ratio and in particular to the optical transmission properties of deep grooves or holes of subwavelength widths. Finally, we describe the behaviour of plasmons in metal nanoparticles and their arrangements, and in continuous metal stripes and wires. We conclude by discussing the prospects opened up by the latest theoretical and experimental progress.

### 16.1.2 What Is a Plasmon?

By analogy with the fluid part of the blood, Langmuir used the word ‘plasma’ to describe the ionised gas state he observed in certain regions of a tungsten-filament light bulb, whose lifetime he hoped to extend [1]. Today the term is used to describe any mixture of mobile, electrically charged particles with varying densities and energies. For the present purposes, we shall take it to mean an electrically neutral ensemble of highly mobile charges (electrons) in the presence of almost fixed charges of opposite sign (ions).

This system turns out to provide a good model for a great many properties of solid bodies, and in particular, metals. For example, the spatial separation of charges of opposite sign generates an electric field, even at long range, which is capable of maintaining collective longitudinal oscillations.

For such an oscillation of the charge density fluctuations to be an eigenmode of the plasma, it must be able to exist without the help of any externally applied field, i.e., the complex dielectric function  $\varepsilon_j = \varepsilon'_j + i\varepsilon''_j$  of the medium

$j$  must be zero:  $\varepsilon'_j = 0$  and  $\varepsilon''_j \ll 1$ . For  $N$  free electrons of charge  $e$  and mass  $m$ , globally displaced relative to  $N$  ions, this so-called bulk plasmon mode will have frequency

$$\omega_p = \left( \frac{4\pi N e^2}{\varepsilon_0 m} \right)^{1/2},$$

and the dielectric function (relative permittivity) of the plasma will be

$$\varepsilon_j(\omega) = 1 - \frac{\omega_p^2}{\omega^2}.$$

In real metals, more localised but nevertheless polarisable electrons can partly screen Coulomb interactions between charge density fluctuations and significantly reduce the plasma frequency.

Charge oscillations are easily excited by an electron beam and detected through the electron energy losses they cause. However, their longitudinal character means that they will not be able to couple with a transverse electromagnetic wave such as light. Hence, the resonant response of an isotropic metal observed around  $\omega_p$  under the effects of light excitation arises purely from the fact that, at the long wavelengths of optical experiments, the transverse and longitudinal dielectric functions assume the same values. Even close to  $\omega_p$ , there is not therefore any direct coupling between photons and bulk plasmons in the metal during transmission or reflection of light beams.

Since the work of Maxwell, we know that, in passing from a medium 1 to a medium 2, the component of the electrical displacement vector  $\mathbf{D}$  perpendicular to the interface will be conserved. Furthermore, in the absence of external charges, the perpendicular component of the electric field is symmetrical with respect to the interface where the polarisation charges generating it are located. As for bulk plasmons, oscillations in the fluctuations of this surface charge density occur in the interfacial plane.

For non-magnetic materials ( $\mathbf{D} = \varepsilon_j \mathbf{E}$ ), these oscillations will occur without external field whenever the above two conditions become compatible, i.e., whenever  $\varepsilon_2 = -\varepsilon_1$ . If medium 1 is a dielectric with positive permittivity, this condition will be satisfied by a medium 2 whose dielectric function becomes negative at certain frequencies, which is what happens in ionic crystals in the so-called Reststrahlen region ( $\omega_{\text{TO}} < \omega < \omega_{\text{LO}}$ ), or again in the Drude metals at low frequencies. In the latter case, if medium 2 can be described by a plasma such that

$$\varepsilon_2(\omega) = 1 - \frac{\omega_p^2}{\omega^2},$$

the interface modes will appear at the frequency

$$\omega_i = \frac{\omega_p}{(\varepsilon_1 + 1)^{1/2}}.$$

If in addition medium 1 is the vacuum, we obtain the plasmon surface frequency

$$\omega_s = \frac{\omega_p}{\sqrt{2}}.$$

For such free metal surfaces, these compressive waves of surface charge density will be localised in a thickness of the order of 0.1 nm.

In the case of a metal particle with permittivity  $\varepsilon_2$ , immersed in a dielectric medium with function  $\varepsilon_1$ , the resonant frequency of the plasmon will depend on the geometry of the particle. When the latter can be described by a depolarisation factor  $L$ , e.g.,  $1/3$  for a sphere, and up to three distinct values lying between 0 and 1 for an arbitrary ellipsoid, the resonance condition along each of the principal axes is given by  $\varepsilon_2 = \varepsilon_1(L - 1)/L$ . For a spherical particle, this relation becomes  $\varepsilon_2 = -2\varepsilon_1$ . Moreover, if it is also metallic,  $\omega_r = \omega_p/(2\varepsilon_1 + 1)^{1/2}$ , and  $\omega_r = \omega_p/\sqrt{3}$  for  $\varepsilon_1 = 1$  (see below).

### Plasmon Dispersion Relations

Collective longitudinal oscillations in the electric charge density resonate at a specific frequency where the dielectric function of the medium vanishes, i.e.,  $\varepsilon'_2(\omega_p) = 0$ . This is the bulk plasmon, illustrated top left in Fig. 16.1, whose frequency  $\omega$  (or energy) depends weakly on the wave vector  $k_x$ , as can be seen from the dispersion curve labelled (1) in the figure. For an interface in the  $x, y$  plane separating media 1 and 2 (lower diagram), the resonance condition for the interface plasmon is  $\varepsilon'_2(\omega_r) = -\varepsilon'_1(\omega_r)$ .

In the case of an ideal Drude metal, whose dielectric function  $\varepsilon_2(\omega)$  is shown on the left of Fig. 16.1, one obtains the dispersion curves labelled (2) and (3) depending on whether the medium with permittivity  $\varepsilon_1$  is air or another dielectric, respectively. When the component  $k_x$  of the wave vector parallel to the surface increases, this wave changes from being photonic to plasmonic.

For a spherical metal particle like the one shown upper right in Fig. 16.1, resonance occurs when  $\varepsilon'_2(\omega_r) = -2\varepsilon'_1(\omega_r)$ , i.e., when  $\omega_r = \omega_p/\sqrt{3}$  if  $\varepsilon_1 = 1$  [curve (4)]. Although this mode cannot propagate (localised mode), it can couple with light when curve (4) crosses the straight line  $\omega = ck_x$  corresponding to light [curve (5)], in contrast to the non-radiative waves described by the other dispersion curves, which do not intersect this straight line.

#### 16.1.3 Dispersion Relations, Coupling with Light, and Applications

The frequency  $\omega$  of a bulk plasma with wavelength  $\lambda$  depends on the wave vector  $k_x = 2\pi/\lambda$  via a quadratic dispersion relation of type  $\omega^2 = \omega_p^2 + Ck_x^2$ . To account for dispersion in the interface modes already described, one must introduce the mixed longitudinal and transverse electric field associated with coherent longitudinal fluctuations of the surface charge in the plane  $z = 0$ , given by

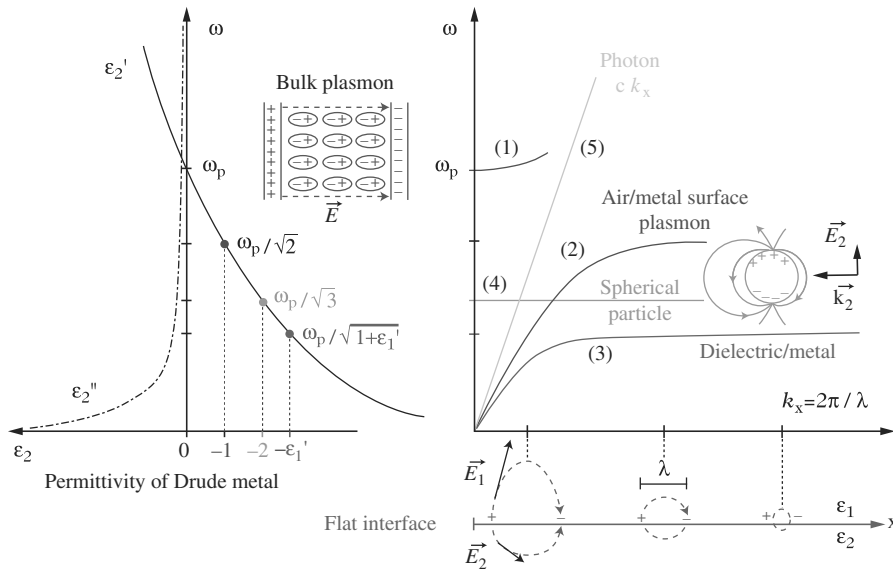


$$E = E_0^\pm \exp \left[ i(k_x x \pm k_z z - \omega t) \right].$$

The associated magnetic induction is perpendicular to the direction of propagation and parallel to the plane of the interface (TM mode).

It can be shown [2] using Maxwell's equations that the wave vector  $k_x$  is continuous across the interface, that it is related to the normal components  $k_{zj}$  ( $j = 1$  or  $2$  depending on the medium) by  $k_x^2 + k_{zj}^2 = \epsilon_j(\omega/c)^2$ , and that the latter satisfy  $k_{z1}/\epsilon_1 + k_{z2}/\epsilon_2 = 0$ . This leads to an expression for the dispersion relation for the surface plasmon:  $k_x = (\epsilon_1\epsilon_2/\epsilon_1 + \epsilon_2)^{1/2}\omega/c$ , which shows (see Fig. 16.1) that, at a given frequency, the wave vector of the surface plasmon is always greater than the wave vector of light in the dielectric. This is a non-radiative wave.

In order to get the surface plasmon to couple with light, one must first increase its wave vector, e.g., by illuminating the surface of the metal by an evanescent TM wave escaping from a wave guide (total reflection prism or fibre), or again, using a surface with well-adjusted and periodic roughness, such as a metal grating. In the two cases described below, the new dispersion curve of the light crosses that of the surface plasmon and an energy transfer can take place between the two modes. A minimum of reflected intensity is



**Fig. 16.1.** Plasmon frequency in different environments and for different shapes as a function of the real part  $\epsilon_2'$  and imaginary part  $\epsilon_2''$  of the complex dielectric function of the metal (left) and the parallel wave vector  $k_x$  at the interface (right). Dashed curves in the bottom image represent electric field lines associated with  $k_x$  on either side of the interface

then observed. Pointlike surface features or random roughness also give rise to coupling with varying degrees of localisation.

In practice, the surface plasmon exists whenever  $\varepsilon'_2 < -\varepsilon_1$ , and the resonance will become more pronounced as  $\varepsilon''_2$  is made smaller, which favours silver, gold, or indium. The propagation length of surface plasmons along the interface is given approximately by

$$\delta = \frac{c}{\omega} \left( \frac{\varepsilon_1 + \varepsilon'_2}{\varepsilon_1 \varepsilon'_2} \right)^{3/2} \frac{\varepsilon_2''}{\varepsilon_2''}.$$

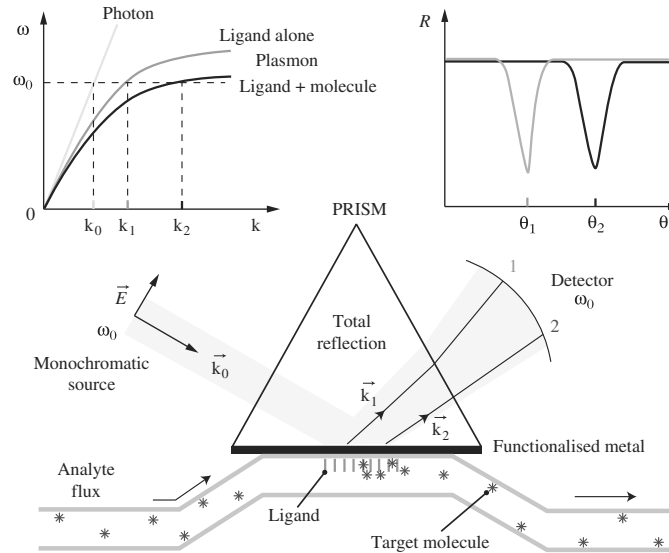
This length can reach values of almost 1 mm for rare metals in the near infrared, but generally remains small (around 10  $\mu\text{m}$ ) in the visible. Moreover, it can be shown that  $k_{zj}$  becomes imaginary if  $\varepsilon_1 + \varepsilon_2 < 0$ , which explains the evanescent nature of the field in directions normal to the free surface of a metal, and also the confinement of the field at the interface. The wave extends into the metal by a distance that is several orders of magnitude less than what would be expected in the dielectric, where almost all the electromagnetic field is concentrated, amplified with respect to the incident wave by a factor which varies at resonance from a few units to a few tens depending on the nature of the metal [2]. Combined with the large variation in the frequency of the interface plasmon with the dielectric constant of the dielectric, these properties explain why surface plasmons are so sensitive to chemical species present on the surface.

The potential of surface plasmons for detecting gases and biological molecules was demonstrated as long as 20 years ago [3]. Today, this technology has been commercialised by several manufacturers and it occupies a dominating position for direct observation of biomolecular interactions in real time and without molecular tagging, detections being made in the zone where the surface wave is coupled with light [4].

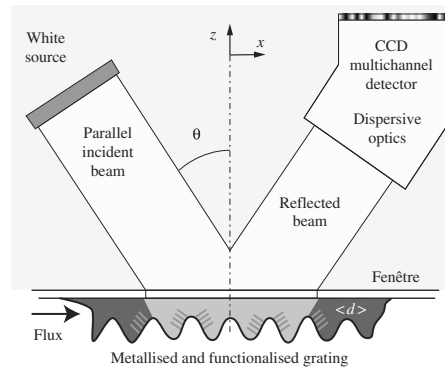
## Surface Plasmon Sensors for Biochemistry

The presence of molecules on the surface of a thin metal film alters the dispersion curve  $\omega(k)$  of surface plasmons. The Kretschmann prism shown schematically in Fig. 16.2 couples these surface waves with TM polarised light for direct detection of biological molecules. In this configuration, the light wave with wave vector  $k_0$  in the prism is totally reflected at the interface between a prism and a metal film of thickness about 50 nm, exciting a surface plasmon on the outer face which is exposed to the medium under analysis. As can be seen from Fig. 16.2, coupling becomes possible between the plasmon and the evanescent wave when  $k_0 \sin \theta = k_{\text{sp}}$ . For a given interface, this happens at a well defined wavelength when the detection angle is fixed, or conversely, at a well defined detection angle when the wavelength is fixed.

Likewise, if an interface between a metal and a dielectric is perturbed in a periodic way in one space dimension (see the grating in Fig. 16.3), the incident light wave is diffracted in several directions. As shown in Fig. 16.4, the parallel component of the wave vector of these diffracted beams differs from that of the incident beam by



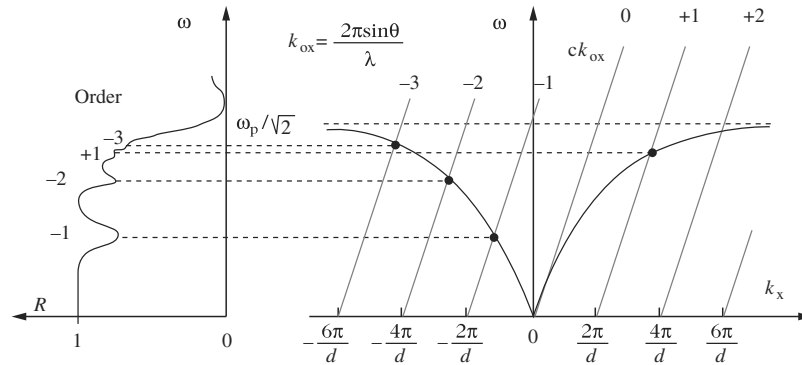
**Fig. 16.2.** Optical sensor using a prism to detect chemical species in solution, showing the angular detection of a change in optical index near a metal surface at the plasmon resonance frequency. The dark angular region of the reflected beam corresponds to losses due to coupling with surface plasmons. Biochemical specificity is ensured by the appropriate ligands



**Fig. 16.3.** Optical sensor using a metallic or metallised grating to detect chemical species by wavelength. Biochemical specificity is ensured by functionalising the metal surface with appropriate ligands

a multiple of the grating vector and can therefore coincide with that of the surface plasmon. This gives rise to various resonances at angles and wavelengths that will change if the dispersion curve of the plasmon is modified.

For angular interrogation, prism and grating devices have very similar theoretical sensitivities which depend only slightly on the nature of the metal, but very strongly



**Fig. 16.4.** *Left:* Reflectivity spectrum of a metal grating with period  $d$  for a given angle of incidence. *Right:* Dispersion curve of the plasmon associated with this surface. The coupling between plasmon and light made possible by the periodicity of the grating gives rise to losses which appear at certain frequencies as minima in the reflectivity. The dispersion curve and the various coupling frequencies depend sensitively on the environment of the surface. This is the basic principle underlying the wavelength detection of chemical species

on the wavelength and detection angle. For spectral interrogation, the prism system is an order of magnitude more sensitive than the grating device. Even if they depend heavily on the characteristics of the metal layer, ATR or guided geometries do not require the same transparency in the analysed medium as grating devices do. This extends the usable wavelength range. Most non-miniaturised commercial systems use the ATR geometry with angular interrogation, but a trend towards analysing modulated signals, and especially towards a higher level of miniaturisation, has laid this choice open to question.

As can be seen from the above discussion, this coupling is achieved by attenuated total reflection (ATR) arrangements which slow down the light in a prism or in a planar wave guide, or else using the properties of light diffracted by a periodic grating. Although it is sometimes the reflected light intensity at resonance that is measured, it is more common to evaluate an angular deviation or a change in wavelength by making a multichannel optical detection, which amounts to measuring the parallel wave vector or the light energy, respectively, at the plasmon resonance.

The dielectric is chosen in accordance with the application. For example, a porous thin layer is appropriate for detecting humidity, whereas a film with high thermo-optical coefficient can be used to measure temperature variations. The metal surface can also be functionalised by radicals so that the dispersion curve of the plasmons is affected in a specific way by local index variations resulting from targeted chemical associations or disassociations. The sensitivity of these devices can reach  $1 \text{ pg/mm}^2$ , which is not quite sufficient for

detecting low concentrations of light molecules, but is perfectly adequate for many biological applications.

In order to exploit the intrinsic advantages of optical fibres, wave guide coupling has also been perfected [5] and even commercialised. The core of the fibre is stripped bare and metallised with a gold or silver layer. One can choose between spectral interrogation using a multimode fibre, or measuring the intensity, normalised to an integrated reference, at fixed wavelength, using a single-mode fibre, which significantly reduces the cost and bulk of the apparatus. These fibre-optic plasmon sensors prefigure other technical advances in which optics, microfluidics and signal processing are integrated into a multichannel device.

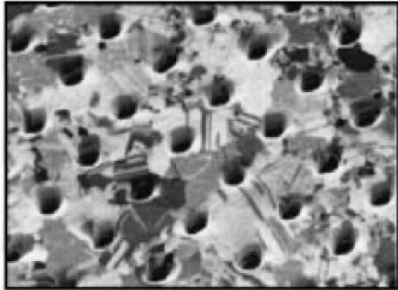
#### 16.1.4 Optical Transmission Through Subwavelength Apertures

When an opaque film made from a material with positive permittivity is pierced in a periodic manner with significantly subwavelength holes and the resulting screen is illuminated by a light beam, the intensity of the transmitted beam is zero in the far field owing to diffraction effects. However, if the opaque film is metallic, part of the incident light is transmitted, as shown in Figs. 16.5 and 16.6. This recent observation [6], not predicted by theory, illustrates in a quite spectacular way the manner in which surface plasmons of a metal can allow light signals to circumvent the Rayleigh criterion. Indeed, the wavelengths of the transmission peaks depend on the symmetry and period of the array of holes, as well as the incidence angle, in a way which indicates that the surface plasmons of the metal play a crucial role in the transmission process.

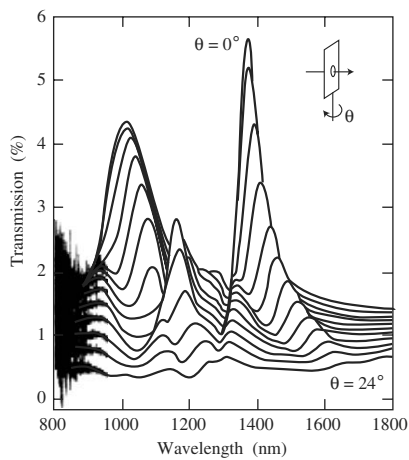
Figures 16.7 and 16.8 show that a similar result is obtained with a metallised silica grating mounted on a silicon wafer and etched with slits or grooves of rectangular cross-section. In this less symmetrical case, where the depth of the grooves is greater than their width, the propagating nature of the electromagnetic modes of the apertures is enough to transmit the localised energy through to the surface and in front of the slits by the excitation of plasmon or cavity modes [7]. When these two modes are present at the same time, particularly large local amplifications are observed in the field.

This scenario indicates that an isolated slit or hole can also transmit certain wavelengths much longer than its own spatial dimensions. Indeed, this has effectively been observed in the microwave and visible range, with energy concentration and directive emission effects provided that the surfaces comprise suitable periodic structures in the vicinity of the aperture. Still concerning single apertures, a saturation phenomenon has been observed in the visible at high photon fluxes [8].

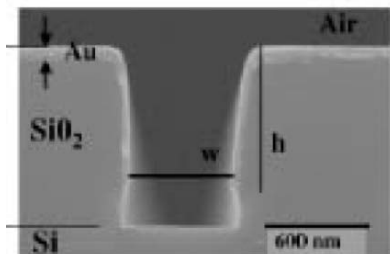
Although the quantitative interpretation of these results is still the subject of some debate, they clearly open up new possibilities in a great many different areas, such as near-field optical microscopy, deep UV photolithography, and optical multiplexing, or even quantum data processing, since an entangled



**Fig. 16.5.** Square array (period 900 nm) of 150-nm diameter holes pierced through a self-supporting silver film of thickness 200 nm. Courtesy of the American Physical Society [10]



**Fig. 16.6.** Zero order optical transmission for angles of incidence between 0 and 24° in steps of 2°. Spectra are shifted vertically by 0.1% for greater clarity. Courtesy of Macmillan Magazines [11]

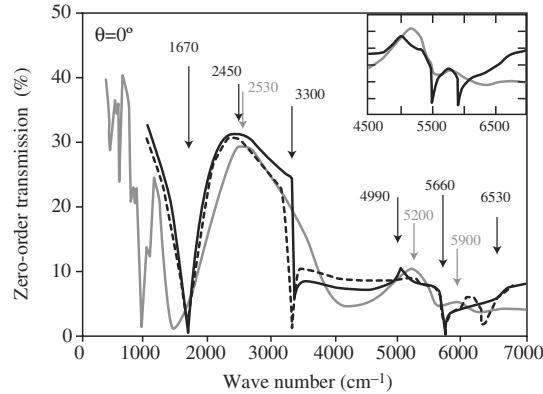


**Fig. 16.7.** Scanning micrograph of a grating with almost rectangular profile, made from 1 100-nm silica strips on silicon, metallised at the top and on the side walls by oblique angle vapour deposition. There is no metal deposited on the silicon surface at the bottom of the groove. Courtesy of the American Physical Society [7]

photon pair conserves its correlation when it passes through a metallic sieve in which the holes have subwavelength dimensions [9].

### Photon Sieves

The polycrystalline silver film of thickness 200 nm shown in Fig. 16.5 has been pierced by a focussed ion beam (FIB) to produce a square array (period 900 nm) of holes, each measuring 150 nm in diameter. At certain wavelengths which depend on the angle of incidence, everything happens as though a fraction of the incident photons



**Fig. 16.8.** Experimental zero order transmission spectrum (*grey curve*) of the grating shown in Fig. 16.7, for normal incidence (except for the *insert*, which is the result for incidence slightly offset from normal) compared with numerical calculations in modal approximation for one mode (*dashed curve*) or four modes (*thick black curve*). *Arrows* indicate wave numbers of various resonances predicted for this system by calculations. Courtesy of the American Physical Society [7]

pass through this nanoscale sieve, as shown in Fig. 16.6 for the wavelength range 800–1800 nm. Another example of transmission by subwavelength apertures is provided by the metallised grating with period  $1.75\ \mu\text{m}$  and rectangular cross-section imaged in Fig. 16.7. This was fabricated by selective dry etching of a silica layer of thickness  $1\ 100\ \text{nm}$  on a silicon wafer, followed by vapour deposition of  $5\ \text{nm}$  of Ti and  $60\ \text{nm}$  of Au at an oblique angle of incidence so as to avoid coating the bottom of the grooves. Here, too, transmission peaks observed between  $6\ 600$  and  $1\ 000\ \text{cm}^{-1}$  ( $1.5$ – $10\ \mu\text{m}$ ) through slits of width  $600\ \text{nm}$  depend on the angle of incidence. Numerical simulations in the modal approximation (Fig. 16.8, black dashes) show that the transmission peaks observed at normal incidence (Fig. 16.8, grey curve) can be attributed to the rear metal/silicon [SP(Si)] interface plasmon, the front metal/air [SP(air)] interface plasmon, or the first cavity mode (CM). As can be seen from the insert of Fig. 16.8, the calculated peak splits into two if the angle of incidence is moved just a few degrees from the normal, which corresponds to experimental collimation.

### 16.1.5 Metal Nanoparticles

We have seen that the plasmon frequency of an isolated nanoparticle depends mainly on its shape and the dielectric properties of the material composing it and the material composing the surrounding medium. The formulas given above, which assume that the particle has a well defined area, do not apply to clusters with sub-nanoscale dimensions, which are bulk polarised and will not be discussed here, nor to isolated particles with sizes above about 5% of the wavelength.

For assemblies of spheres with sizes greater than this limit or in which the spacing is reduced, the optical characteristics are modified in a complex but well understood way [12,13]. They are illustrated just as well by the colours of Roman pigments and medieval stained glass as by the spectral response of a periodic arrangement of nanoparticles. The latter can be evaporated through a flat mask of latex nanospheres, or else defined by electron beam lithography. A suitable periodic spatial arrangement can even be used to amplify or inhibit resonances in certain wavelength bands, as in photonic band gap structures (see Sect. 16.3) [14].

One particularity of plasmons localised on particles is the very strong electromagnetic field near their surface. More precisely, at a point  $M$  located at a distance  $d$  from the center of a metal sphere of radius  $r$ , the electric field  $E_M$  which results from superposing the incident field  $E_0$  and the dipole field induced in the metal sphere will be amplified by a factor

$$A(\omega) = \frac{E_M}{E_0} \propto \left( \frac{r}{r+d} \right)^3 \frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_2 + 2\varepsilon_1} .$$

Near the plasmon frequency, this factor, which arises with a power of 4 in the amplification of surface-enhanced Raman scattering (SERS), has been used to carry out ultrasensitive chemical and structural analyses by Raman microspectrometry on various untagged biological systems, including single molecules placed on a suitable substrate [15,16]. According to observations by apertureless scanning near-field optical microscopy (SNOM) (scattering mode), the local amplification of the field is still more pronounced in spaces separating metal nanoparticles when they are very close together (10 nm) [17]. Such configurations are also found in coatings by metallic islands used as active substrates in SERS or as optical fibre sensors with enhanced detectivity [18].

Other SNOM systems using total reflection geometries can also be used to visualise, with submicron resolution, the near field of a single metal particle prepared by nanomanipulation or electron beam lithography (see Figs. 16.9 and 16.10). Simulations confirm that the two relatively strong and extensive lobes thus detected under resonant excitation and located in the horizontal plane on either side of the particle can be identified with the near field of a vertical dipole. In addition, they confirm that the stationary wave also observed is due to interference between the incident and scattered fields [19].

The damping of plasma oscillations in a single particle can be measured via the excitation spectroscopy of such SNOM images. This then allows one to find the lifetime of these excitations. Values of the order of 10 fs are obtained, in agreement with time decay measurements on assemblies of particles [20].

These considerations lead one to pose the following question: How can this spectral selectivity, combined with the ultrafast character and nanometric size of the particles, be exploited to achieve individual addressing of nanoscale components?

A first hint at an answer is provided by the properties of a periodic chain of nanoscale dots, where the near field is almost totally confined to the spaces



separating the dots and where the effective transport of an electromagnetic signal can reach several microns (see Fig. 16.10). Numerical simulation of this nanostructure (see Fig. 16.11) reveals that it is the coupling between individual nanoparticles that produces the density of states of a collective oscillation, whose near field is much more localised laterally (down to a tenth of the wavelength) than it would be around an isolated particle (see Fig. 16.9). Like the near field of a continuous wire, the near field in this case depends sensitively on the polarisation of the incident light, in contrast to the depolarised response which characterises pointlike nanostructures.

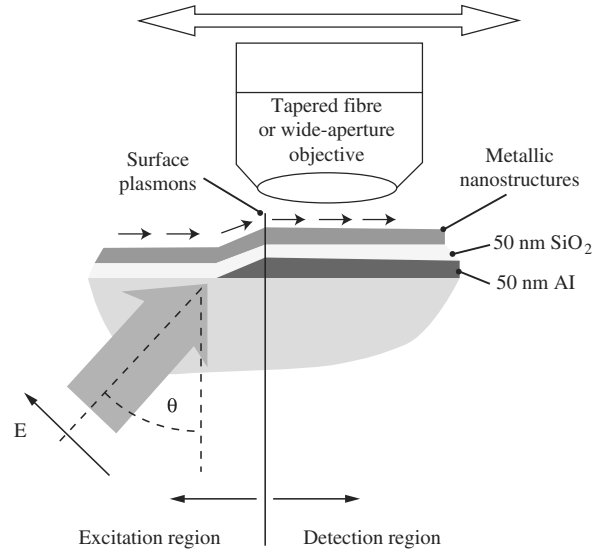
However, if a metal nanoparticle is placed a few tens of nanometers away from a nanowire and excited in such a way (with regard to wavelength and polarisation) that the isolated nanowire is not resonant, a clear strengthening of the near field is observed around the nanowire [19]. Other cases of efficient guided propagation and coupling have been observed for a nanowire connected to a metallic stripe and for a nanoparticle placed just beyond the triangular end of a microscale metallic stripe [19].

Whilst the propagation length appears to be limited to a few microns [20] by radiative losses when the width of the flat metallic wave guides reaches submicron levels (see Fig. 16.12), it would seem that it can actually remain so for wave guides with nanoscale cross-section. In particular, for bimetallic rods with length  $10\mu\text{m}$  and diameter  $10\text{nm}$  comprising a clear Au/Ag interface, unidirectional propagation of excited plasmons is observed in the near infrared [21].

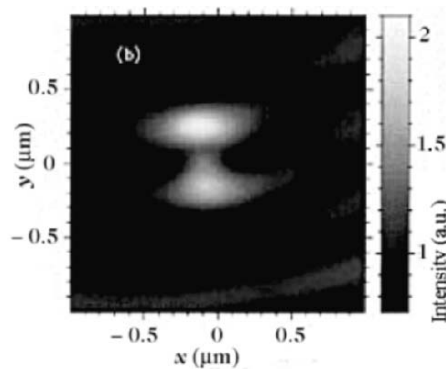
These examples show that several elementary operations of optical addressing can be achieved using metallic nanostructures. For wavelengths greater than  $10\mu\text{m}$  (far infrared), quantum cascade lasers and THz sources already use surface plasmon wave guides confined to the metal/semiconductor interface, with modes controlled by patterning the metallic layer. As in distributed feedback lasers, a two-metal grating made from alternating microscale stripes of two different metals can be fabricated in such a way as to ensure the single-mode nature of the wave guide [22]. Likewise, with the size reductions that will be required by future integrated elements, it seems likely that the rapid damping of surface plasmons will no longer be critical at the infrared wavelengths used in telecommunications and that integrated plasmonic devices, first flat but eventually 3D, will come into existence.

### Plasmon Observation and Guiding

The surface plasmons of a film or metallic nanostructures are easily excited by total internal reflection through a prism coated with a transparent film. As can be seen from Fig. 16.9, the presence of surface plasmons can then be detected by placing a tapered fibre in their near field (evanescent wave). This frustrates total reflection and a signal proportional to the optical near field of the surface can be picked up. For a square gold particle of side  $100\text{nm}$  and height  $40\text{nm}$ , the two lobes predicted by theoretical calculations on either side of the particle, oriented by the direction of



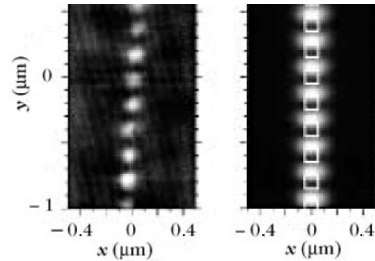
**Fig. 16.9.** Local detection by fibre or objective of light signals associated with plasmons in metallic nanostructures in a region (*right*) where they are not excited by total reflection (obturation by an opaque aluminium film)



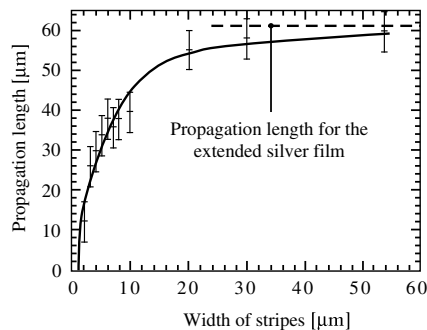
**Fig. 16.10.** Near-field optical image of a gold particle ( $100 \times 100 \times 40 \text{ nm}^3$ ) placed slightly to the left of the origin of the spatial coordinates. Courtesy of the American Physical Society [19]

propagation of the exciting wave, can be observed at resonance (see Fig. 16.10). If such particles are now lined up to form a chain with spacing 100 nm, a spectacular confinement or squeezing of the electromagnetic field between the nanoparticles is observed at the same wavelength (see Fig. 16.11). This effect is clearly reproduced by theoretical calculations.

In order to make a clear separation between the excitation region and the detection region, an optically opaque strip of Al can also be deposited on the silica before the metallic nanostructures (see Fig. 16.9). If the latter are long stripes of thickness



**Fig. 16.11.** Comparison between the near-field optical image of a chain of gold particles and its numerical simulation, showing the quality of the calculations. *Bright regions* indicating strong fields are located in the 100-nm interval separating the particles. Courtesy of the American Physical Society [19]



**Fig. 16.12.** Propagation length of surface plasmons determined experimentally at 633 nm for different widths of silver stripes of height 70 nm. Courtesy of the American Institute of Physics [20]

70 nm and widths varying over 1–50  $\mu\text{m}$ , the propagation length of surface plasmons in such wave guides can be determined by detecting the light resulting from their scattering by the corrugations in the far field using a large-aperture objective. The significant effect of reducing their dimensions can then be confirmed for silver and in the visible, as shown in Fig. 16.12.

### 16.1.6 How Far Can Plasmons Take Us?

There will be many openings for the integrated subwavelength optics that surface plasmons promise to make possible, including applications in areas that are largely ignored today [23]. The significant enhancement of the electromagnetic field in the vicinity of highly curved metal surfaces will be put to use in detection devices with ever-increasing spatial resolution, whether it be in SERS, integrated sensors, composite materials with negative refraction [24], or near-field optical microscopy. Moreover, metal plasmon particles and their immediate surroundings are the scene of an extraordinary concentration of electromagnetic energy, mainly limited by radiative losses. This situation should favour the appearance of many nonlinear phenomena, including the amplification of surface plasmons by stimulated emission of radiation (spaser) predicted in certain theoretical approaches [25].

## 16.2 Semiconductor Quantum Dots

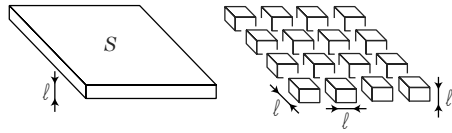
The reduction in the number of dimensions of objects made from semiconducting materials has been pursued for thirty years now. Indeed, research soon demonstrated that the reduction in size of a sample in 1, 2 or 3 directions would reduce the number of degrees of freedom open to electrons, thereby changing the number of fundamental parameters of the material: band gap, electron density of states, effective mass, etc. Semiconducting materials were no longer intangible substances, but could become the subject of engineering, to adapt them to some given function. The technological difficulties involved in fabricating 2D materials called quantum wells were quite rapidly overcome. The same cannot be said for 1D materials (quantum wires) or 0D materials (quantum dots). Indeed, in order for these entities to acquire significantly different properties, their dimensions had to be very accurately controlled, especially as size fluctuations assume a growing relative importance when the size decreases. It is through the successful control of self-organisation properties that it has been possible to overcome these difficulties and make components from semiconductor quantum dots.

### 16.2.1 Semiconductor Lasers: From Quantum Wells to Quantum Dots

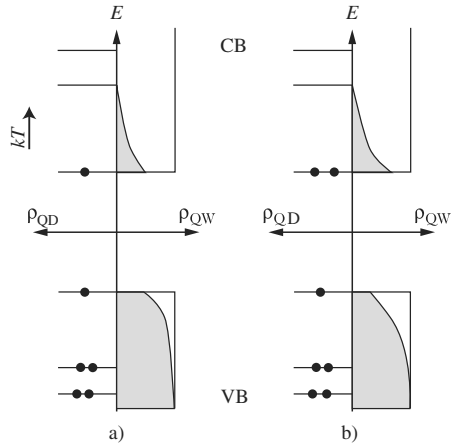
During the 1980s, the performance of semiconductor lasers was considerably improved with regard to threshold current, efficiency, and maximum modulation frequency, by using an active quantum well material. The quantum well laser is an economical, reliable, high performance component which was commercialised on a large scale after a few years of development. For example, it can be found today in a low cost version in all compact disk players, and in a more elaborate form in high speed fibre optic transmission systems. In 1982, Arakawa and Sakaki suggested that a reduction in the dimensionality caused by structuring the active medium in quantum wires (1D) or in quantum dots (0D) would generate a new technological breakthrough. This prediction was without doubt the main motivation for the many studies subsequently devoted to quantum dots. The development of self-organised growth techniques led to the design of quantum dot lasers in 1994, but ten years on, these have still not found a niche in the market.

After a brief summary of the potential benefits of quantum dot lasers, we present the principal optical properties of semiconductor quantum dots. We then identify the main difficulties, explaining why this predicted revolution never actually took place. On the other hand, in optoelectronics, there are in fact several important technological niches for quantum dots which we shall examine to end this section.

The potential use of quantum dots for lasers as conceived at the beginning of the 1980s is illustrated in Fig. 16.14, which shows schematically how the gain can be built in a 2D system and in an ideal 0D system comprising an ensemble



**Fig. 16.13.** The problem of the quantum dot in the 1980s: can the semiconductor laser with separate confinement shown above be improved by replacing the quantum well by a plane filled with quantum dots?



**Fig. 16.14.** Conduction and valence densities of states  $\rho$  for a quantum well and a plane of ideal quantum dots, assumed to have the same gap. The filling of these states is represented at the inversion threshold (a) and above threshold (b). Occupied states are indicated by shaded regions and black dots

of identical dots. This is the strong confinement regime, for which the distance between the quantum levels of the dot is large compared with the heat energy  $kT$ . With the quantum well, the injected electrons (holes) spread out across the conduction (valence) miniband according to the Fermi–Dirac distribution. When the inversion condition is satisfied, the quantum well exhibits a gain. The laser effect occurs when this gain is enough to compensate the optical losses of the laser cavity, at the gap energy  $E_g$  of the quantum well, the energy for which the population inversion and the gain reach a maximum. Note that, to reach the laser threshold, the quantum states located within  $kT$  from the band edge must be significantly populated. These injected carriers, which do not contribute directly to the gain at  $E_g$ , can nevertheless recombine by spontaneous emission. One must therefore inject a minimal current, the threshold current, to maintain the quantum well in this non-equilibrium state. Naturally, there may be other recombination channels, of a non-radiative nature, which will tend to increase the value of this threshold current, but in a ‘good’ quantum well laser, spontaneous emission will dominate.

It is immediately clear that the 0D system has many potential benefits. Just one electron–hole pair per dot is enough to reach the inversion threshold, and all injected carriers contribute to the gain. The fact that ineffectual electron states are not populated can potentially reduce the threshold current of the laser by a significant margin. In 2D, when an extra electron–hole pair is injected, the additional gain is spread across a broad spectral band. In 0D, one expects to modulate the gain  $g$  more efficiently by playing on the carrier

density  $n$ . This higher differential gain  $dg/dn$  is potentially useful for increasing the maximal modulation frequency of the laser and improving its spectral quality above threshold.

Provided that the interlevel spacing remains much greater than  $kT$ , the populations of the quantum levels do not change. One thus expects the threshold current of quantum dot lasers to be insensitive to temperature, which would be an extremely useful property in practical terms. It is therefore easy to understand why this set of predictions generated such a surge of interest in quantum dots.

In a real system, there is a fluctuation in the sizes of quantum dots, which manifests itself through a variation in the carrier confinement and gap energies from one dot to another. Provided that the broadening  $\Delta E_{\text{QD}}$  of the gain curve remains lower than  $kT$ , one can hope to obtain better performance from the QD laser in terms of threshold current and differential gain. On the other hand, it is easy to see that the technological difficulties associated with this goal are colossal. In order to achieve the strong confinement regime for both electrons and holes, a typical size for the quantum dots would have to be less than about 10 nm in each space direction. For such a small quantum system, relative size fluctuations must then be of the order of 5–10% at the most, if the replacement of the quantum well by a plane of dots is to be of any advantage.

These tough requirements explain why the first attempts at nanofabrication of quantum dots, based on a combination of electron beam lithography and reactive ion etching, were abandoned at the beginning of the 1990s, although when optimised this process was able to produce quantum dots with satisfactory emission efficiencies [26]. Indeed, on a standard resist of PMMA type, electron beam lithography can reach accuracies of the order of 3 nm, which represents a relative size fluctuation of about 30% for a 10-nm QD!

It is remarkable that the technique known as self-organised epitaxial growth described in Chap. 2 [27] is able to spontaneously procure, in the best cases, planes of quantum dots with lower fractional size fluctuations, e.g., of the order of 7% for the InAs/GaAs system. This achievement, combined with other novel benefits such as the collective aspect, cleanliness, and low cost, explain why this fabrication process has come to dominate, spreading to a great many research centers since about 1995.

In general, the condition  $\Delta E_{\text{QD}} < kT$  is far from being satisfied, with the notable exception of the InAs/GaAs system, which is easily the best understood. A line width  $\Delta E_{\text{QD}}$  of the order of 20 meV is observed at 300 K for the biggest InAs dots, hence the least sensitive to size fluctuations, that we are capable of fabricating. Their emission wavelength is somewhere near 1.3  $\mu\text{m}$ . The broadening of the gain curve is naturally accompanied by a reduction in the maximum available gain, since the dots are not all functioning in concert at the same energy. One must therefore test whether the number of QDs is enough to lase. To this end, consider the separated confinement structure shown in Fig. 16.13 and assume that it contains a quantum well. When the width of the wave guide is chosen to maximise the amplitude of the guided

mode at the position of the well, the maximal available gain is something like  $100 \text{ cm}^{-1}$  for an InGaAs well.

Let us now replace the well by a plane of QDs in this wave guide of optimal dimensions. For an allowed optical transition, the modal absorption or the gain do not depend on the dimensionality of the system.<sup>1</sup> It suffices therefore to compare the joint density of states  $\rho'$ , i.e., the number of allowed optical transitions per unit energy, for the well and the plane of dots in order to estimate the maximal modal gain.

For a quantum well, this density of states is constant and  $\rho'_{\text{QW}} = Sm^*/\pi\hbar^2$ , where  $S$  is the area of the relevant active medium and  $m^*$  is the geometric mean of the effective masses in the conduction and valence bands ( $m^* \approx 0.04$  for InGaAs). For the plane of quantum dots, assuming a Gaussian size distribution, we obtain

$$\rho'_{\text{QD}} = \frac{4\gamma S}{\Delta E_{\text{QD}}\sqrt{2\pi}},$$

where  $\gamma$  is the surface density of the plane of dots. For a state-of-the-art plane of InAs dots, we have  $\gamma \approx 2 \times 10^{10} \text{ cm}^{-2}$ ,  $\Delta E_{\text{QD}} \approx 20 \text{ meV}$ , which leads to a maximal modal gain close to  $10 \text{ cm}^{-1}$ , i.e., ten times less than for a quantum well! This simple estimate has been confirmed by experimental measurement of the modal gain.

For the non-specialist reader, it should be mentioned here that many papers in the literature speak of a ‘giant’ gain for the quantum dot device. However, this is not the modal gain, but rather the ‘material’ gain that would be seen by an electromagnetic mode if it were entirely confined within the active material. Such a situation is unrealistic in practice and the only relevant parameter is the modal gain, which is unfortunately rather low for an ensemble of QDs. We note in passing that nobody would have the idea of referring the gain of a Ti:sapphire laser to the volume of active atoms!

In a conventional laser, optical losses are mainly due to the low reflectivity ( $\sim 0.3$ ) of mirrors obtained by cleavage, and are of the order of  $20\text{--}30 \text{ cm}^{-1}$ .

<sup>1</sup> In the quantum well or dot, the wave function of the electron (hole) can be expressed as the product of a Bloch function of the bulk crystal  $u_c$  ( $u_v$ ) and an envelope function  $\phi_c$  ( $\phi_v$ ), which describes the variation of its spatial probability amplitude at scales greater than the details of the crystal lattice. It is easy to show that the absorption or the modal gain are proportional to  $|\langle \phi_c u_c | p | \phi_v u_v \rangle|^2 = |\langle \phi_c | \phi_v \rangle|^2 |\langle u_c | p | u_v \rangle|^2$ . For an allowed transition of a quantum well, the electron and the hole have the same wave vector for their propagation in the plane of the layers, and the overlap of their envelope functions is equal to unity. In quantum dots, the envelope functions associated with the first levels of the electron and the hole also have an overlap close to unity in the typical case. We thus find that the modal gain by allowed optical transitions is given by the matrix element between Bloch wave functions of the bulk material, viz.,  $|\langle u_c | p | u_v \rangle|^2$ . It does not therefore depend on the dimensionality of the heterostructure.

To reach the laser threshold, several planes of QDs must therefore be inserted into the active layer.

Since the first demonstration of an InGaAs QD laser in 1994 [28], the performance of these components has made steady progress in terms of threshold current [29], output power [30], efficiency [31], modulation frequency, or sensitivity to temperature [32], sometimes even approaching the performance of InGaAs quantum well lasers of equivalent wavelength, although never exceeding them. The only counterexample here is the current density at threshold. For structures with low optical losses, e.g., lasers with long cavities, the laser effect can be obtained with a single plane of dots. The number of states to be inverted is then much lower than for a quantum well, and this means that the current density at threshold can be reduced to around  $30 \text{ A/cm}^2$  [29], compared with  $50 \text{ A/cm}^2$  at best for a quantum well laser. However, this record-holding laser has a greater length (5 mm) than the standard quantum well laser (300–500  $\mu\text{m}$ ), whence the total current injected at the laser threshold is greater for the QD laser.

What may we expect for the future in this area of research? To improve the present performance of InAs QD lasers, the surface density of the planes of dots must be increased, whilst at the same time reducing the inhomogeneous width of the gain curve. This is likely to be a delicate task because the growth conditions are well understood and they have already been extensively optimised within this kind of system. Even if it eventually proved possible to elaborate planes of QDs with ideal size uniformity, the benefits would still be rather modest. Indeed, as we shall see below, the homogeneous width of the fundamental optical transition at 300 K is of the order of 10 meV for an isolated quantum dot [33], and of the order of 20 meV for a quantum dot in an operating laser. There is therefore no hope of obtaining a much sharper gain curve for a plane of QDs than for a quantum well.

We may conclude from this discussion that quantum dots do not provide a universal solution for improving semiconductor laser performance across the board. In contrast to the quantum well, which quickly stole the scene as active medium in almost all semiconductor lasers in the 1980s, one should not expect to see a long term or generalised replacement of quantum wells by quantum dots in optoelectronic components. Instead, it seems likely that there will be certain technological niches in which the specific properties of quantum dots may prove useful.

Needless to say, the quantum dot component comes into its own when there is no high performance quantum well laser operating in the same wavelength range. As an example, InAs quantum dots can cover the spectral range 1–1.37  $\mu\text{m}$  [30], whilst InGaAs quantum well lasers are limited to a maximal wavelength of 1.1  $\mu\text{m}$ . InAs/GaAs QD lasers emitting at 1.3  $\mu\text{m}$  have been investigated in great depth thanks to a potential application in telecommunications, where they would replace InP/InGaAsP quantum well lasers. The transition from InP materials used today to GaAs materials might help to solve a certain number of practical problems and lower the cost of these

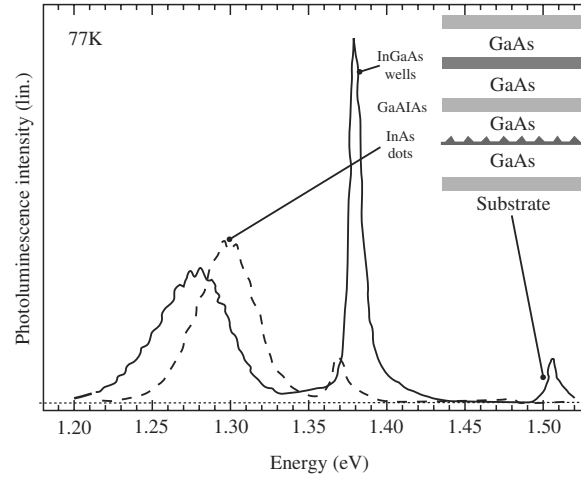


components. More generally, a great deal of effort has been made to extend the spectral range covered by quantum dot lasers by varying the choice of materials used. As an example, the use of quantum dots for sources in the mid-IR (3–5  $\mu\text{m}$ ) would make it possible, by discretising the electron states, to dramatically reduce radiative recombination via the Auger effect, a phenomenon that seriously affects the performance of quantum well components in this spectral range.

A further difference of an elementary but crucial nature that distinguishes quantum wells and dots is connected with the trapping of charge carriers. In contrast to the quantum well, for which carriers remain free to move around in the plane of the layers, the capture of a carrier by a quantum dot allows one to localise it. This effect is particularly useful in light-emitting components, when non-radiative recombination centres such as dislocations or free surfaces are present in the vicinity of the active medium. In this case, carrier diffusion towards these centres can be inhibited by using quantum dots, thereby avoiding their non-radiative recombination.

The first application envisaged was the integration of optoelectronic functions on an Si substrate using epitaxial growth of III–V components on Si. At the time (around 1990), this integration came up against the poor quality of III–V materials epitaxied on Si. Owing to the very high density of dislocations, quantum well lasers exhibited very high threshold currents, with a lifetime of the order of 1 s. It should be remembered that the catastrophic degradation of these lasers was due to a multiplication in the number of dislocations activated by the dissipation of energy generated by non-radiative recombination of electron–hole pairs. If quantum dots could be used instead, this would then provide an efficient and stable alternative. The potential for quantum dots in this context is illustrated in Fig. 16.15, which compares the emission spectrum of a structure based on GaAs, containing a quantum well and a plane of quantum dots, depending on whether it has been epitaxied on Si or on GaAs. The emission of the InAs dots turns out to be the same in both cases, despite the presence of a high density of dislocations ( $10^7 \text{ cm}^{-2}$ ) for growth on Si, whereas the emission efficiency of the quantum well plummets by an order of magnitude. However, this success does not mean that one can make an efficient laser on Si. Indeed, a laser operates in a regime of high electrical injection, for which a significant fraction of the electron population occupies delocalised states of the barrier and remains subject to the phenomenon of non-radiative recombination.

In contrast, carriers can be perfectly localised in a light-emitting diode when it is in use, because the injected current density is typically a hundred times smaller than in a laser. This effect underlies the blue, green and white GaN light-emitting diodes. The commercialisation of these diodes in 1994 was an important event, because the availability of reliable and efficient semiconductor sources in the blue and green greatly increased the scope for applications in the fields of visualisation and lighting. The market for these diodes grew by 30% in 2002 to reach a value of the order of 2 billion dollars per



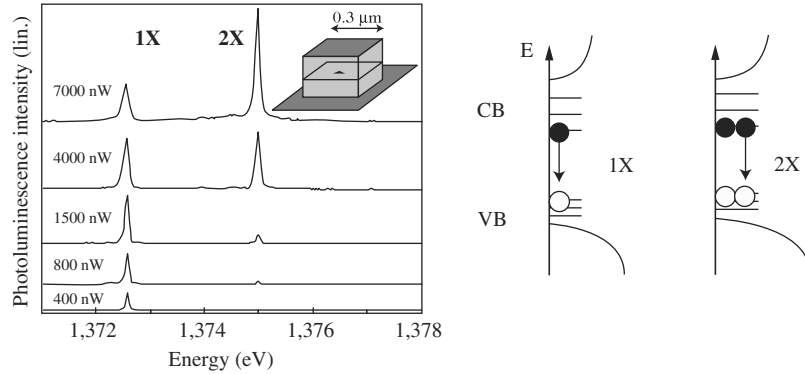
**Fig. 16.15.** Photoluminescence spectra obtained at 77 K for the same structure (shown schematically in the *insert*) grown epitaxially on a GaAs substrate (*continuous curve*) and on a highly dislocated GaAs-on-Si substrate (*dashed curve*) [34]

year. The excellent performance of these components is at first sight rather surprising since no substrate is particularly well suited to their growth. Sapphire or SiC are generally used, but their large difference in lattice parameter compared with GaN leads to very high densities of dislocations, with an interdislocation distance of the order of 1–10  $\mu\text{m}$ . A great deal of research has been carried out to understand the reason for this insensitivity to dislocations. Today it has been clearly demonstrated that the alloy InGaN, used to make quantum wells in the active layer of these components, has a strong tendency to phase-separate [35, 36]. A very dense set of nanoscale inclusions forms spontaneously during the growth of these layers. These inclusions, rich in indium, have a much lower band gap energy than the surrounding material, whereupon they are able to trap charge carriers very effectively.

### 16.2.2 Single Quantum Dots

Other promising applications arise from the highly specific emission properties of single quantum dots. The first problem is to isolate a single quantum dot for optical study. This can be done by defining submicron cavities by electron beam lithography and etching in a sample containing a plane of quantum dots [37]. Figure 16.16 shows a photoluminescence study of a single InAs quantum dot as a function of the power of the exciting optical beam.

For low excitations, its emission spectrum consists of a single line with width much less than  $kT$ . This result, together with very narrow peaks in the absorption spectra, reflects the discrete density of states of the dot and



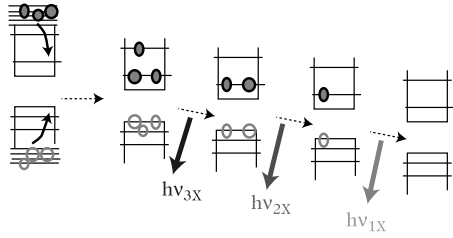
**Fig. 16.16.** *Left:* Photoluminescence spectra at 4 K of a single InAs quantum dot, isolated in an etched GaAs cavity for a range of excitation powers. *Right:* Origin of emission lines 1X and 2X

justifies to some extent the term ‘artificial atom’ that is often used to describe it. However, this artificial atom is not isolated.

When the temperature is raised, the line is observed to broaden. This is due to coupling between this localised electron system and the vibrational modes of the crystal [38,39]. At 300 K, line widths in the range 5–10 meV have been observed by various methods [33]. Another property specific to QDs resides in their ability to store several electron–hole pairs within an extremely small volume, of the order of  $100 \text{ nm}^3$ . Figure 16.16 shows the appearance of a new line when the excitation power is increased [40,41]. This line, whose intensity varies quadratically with the excitation power, corresponds to a situation in which the QD contains two electron–hole pairs. The spectral shift between these two lines results from the Coulomb interaction between the two pairs.

The properties of this rather unusual artificial atom can be exploited to carry out a great many quantum optical experiments usually made with real atoms. Some of the observed effects open the way to the development of original optoelectronic components displaying novel functionalities. In this context solid-state single-photon sources (S3PS) made from quantum dots serve as a good example. Indeed we shall make use of their recent development to illustrate the conceptual renewal afforded in optoelectronics by the association of quantum dots and optical microcavities on the one hand, and by the use of single quantum dots as active media on the other.

A single-photon source is a component that can emit on request light pulses containing one and only one photon. Recall that, for a light pulse emitted by a conventional source, the number  $N$  of photons emitted is poorly defined. This is the case, for example, with a thermal source such as an incandescent light ( $\Delta N \sim \langle N \rangle$ ), or even with a laser source ( $\Delta N = \langle N \rangle^{1/2}$ ). At the present time, the development of single-photon sources is mainly motivated by their use in quantum cryptography. Based on the principles of quantum mechanics,



**Fig. 16.17.** Radiative cascade in a QD and protocol used to generate a single photon [46]

this uses quantum objects to code information, with the aim of guaranteeing absolute confidentiality when information is exchanged [42]. Today, many laboratory experiments are in progress, and even some tests on fibre optic telecommunications networks.

The first protocol put forward coded binary information via the polarisation of single photons and achieved a good compromise between complexity and efficiency. Up to then, a laser source was used, in the absence of any viable single-photon source. This laser source had to be operated at very low power ( $\langle N \rangle \ll 1$ ) to limit the fraction of pulses containing several photons, which might be exploited by a spy. Simple models showed that, by using a true single-photon source, the range of an unconditionally secure link by optical fibre could be increased from 30 to 100 km, or the data flow rate of the link could be increased by more than a factor of 100 at constant range [43].

In the much longer term, an almost perfect single-photon source could be used to calibrate light (and energy) flux, or to develop a quantum computer using single photons as quantum bits (qubits).

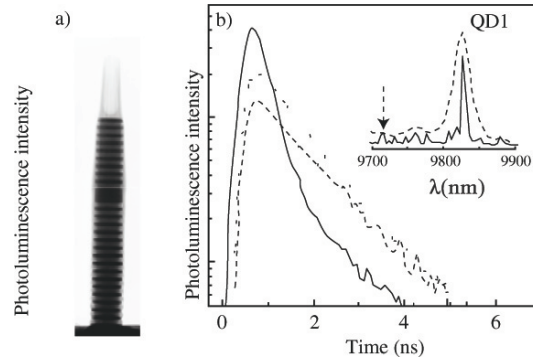
There are two strategies available for making an S3PS:

- Historically, the first suggestion was to use the Coulomb blockade effect. A single photon can be prepared by injecting exactly one electron and one hole into a semiconductor quantum well [44]. However, this approach only gave results at low temperatures ( $T < 0.1$  K) and the presence of metal electrodes in the immediate vicinity of the well seems barely compatible with efficient collection of the emitted photons.
- The second approach consists in setting up a single emitter with discrete electron states. Single-photon emission has recently been demonstrated for a molecule, a colour centre, a semiconductor nanocrystal, and a quantum dot. However, the latter is the only source to have been electrically pumped and to have been able, when inserted within a microcavity, to produce a single-mode S3PS, capable of very efficient generation of single photons prepared in the same spatial mode and with the same polarisation [45]. The latter achievement is particularly useful in practice for quantum cryptography: before information can be encoded by single photons, they must be prepared in a specified initial state. When a non-polarised source is used, half the photons are lost during this preparatory stage.

A QD S3PS requires two things to work correctly: emission of single photons and preparation of these photons in a given mode. Since a QD can simultaneously trap several electron–hole pairs, it is essential to set up a suitable procedure for avoiding the generation of pulses containing several photons. The strong Coulomb interaction between carriers trapped in the QD provides a very simple solution to this problem [46], which has been confirmed experimentally by many groups [45,47,48]. The basic idea is illustrated in Fig. 16.17. An isolated QD is excited by either electrical or optical non-resonant pulse pumping. Several electron–hole pairs, photocreated within the barrier, are quickly captured by the QD ( $\sim 20$  ps). The sequential radiative recombination of these pairs is then observed. Each emitted photon has a specific energy  $X_n$  which depends on the number  $n$  of electron–hole pairs present in the QD when it is emitted. The idea is to carry out spectral filtering of the emission from the QD to prepare light pulses containing a well defined number of photons. By selecting only the line  $h\gamma_{1X}$  of the QD, a pulse containing one photon is prepared for each pump cycle, whilst highly correlated photon pairs can be prepared by filtering the lines  $h\gamma_{1X}$  and  $h\gamma_{2X}$  [49].

By construction, the QDs are contained within a dielectric matrix with high refractive index, so that many of the emitted photons generally remain trapped in the matrix by the phenomenon of total internal reflection. Photons can be more efficiently extracted, and prepared in a common state with regard to spatial mode and polarisation, by integration into an optical microcavity. To date, the most effective way of doing this has been to insert the quantum dot into a pillar microcavity like the one shown in Fig. 16.18a. This microcavity guarantees the 3D optical confinement of a discrete set of photon modes by associating a guiding effect along the axis of the dielectric cylinder with reflection by two distributed mirrors placed on either side of the cavity. However, a micropillar is not an ideal ‘photon box’, because it also contains a continuum of non-confined modes corresponding to photons propagating in such a direction that this guiding and/or reflection do not operate. An emitter placed within this imperfect microcavity will share its emission between these non-resonant modes and, if the spectral matching condition allows it, one or more resonant modes. The problem then is to carry out a single-mode collection of the spontaneous emission.

The Purcell effect provides a very effective solution which can approximate to the ideal regime. The Purcell effect [50] consists in selective enhancement of the spontaneous emission from an emitter in a confined mode M of the micropillar with which it is in resonance. This effect is illustrated in Fig. 16.18b, which shows the time-resolved photoluminescence study of several InAs QDs placed in a micropillar. Note that the radiative lifetime is three times shorter for the quantum dot QD1 which is in resonance with the fundamental mode of the micropillar than for QD2 and QD3 which are only coupled with non-resonant modes. This result shows that, in the case of QD1, spontaneous photons are emitted twice as fast in the discrete mode M as in the ensemble of



**Fig. 16.18.** (a) Transmission electron microscope view of a GaAs/AlAs micropillar (diameter  $1\ \mu\text{m}$ ). (b) Time decay of the emission from three InAs quantum dots placed in the micropillar following excitation by a laser pulse. The quantum dot QD1, in resonance with the mode of the micropillar, displays rapid decay, exemplifying the Purcell effect. Photoluminescence spectra obtained for this micropillar are shown in the *insert*. Under weak excitation (*continuous curve*), the emission of the various QDs broadens significantly, revealing the spectral position of the cavity mode [45]

non-resonant modes, and that a fraction  $\beta = 66\%$  of its spontaneous emission is injected into the mode M by virtue of this dynamic effect.

Further studies carried out at CNRS/LPN in Marcoussis, France, and Stanford University, USA, confirm that a quantum dot in a micropillar already performs sufficiently well to be considered for quantum cryptography. In particular, the probability of emitting a photon is greater than 40% and the probability of emitting several photons is reduced by an order of magnitude compared with an attenuated laser source of the same average intensity. Short term studies in this area are oriented towards the development of a practical ‘plug-and-play’ source for quantum cryptography, combining electrical pumping and single-mode emission at a telecommunications wavelength of  $1.3\ \mu\text{m}$ . On a more fundamental level, research is under way to generate other quantum states of light, such as entangled photon pairs.

On the conceptual level, the S3PS is the first optoelectronic component to operate on the basis of a cavity quantum electrodynamic (CQED) effect, viz., the Purcell effect. During the 1980s, experiments carried out on atoms placed inside optical cavities showed that it was possible to modify the optical properties of an atom to a large extent, in particular its spontaneous emission, by controlling its coupling with electromagnetic radiation [51]. However, it was not until the advent of the QD, playing the role of an artificial atom at low temperatures, that any of these CQED effects, including the Purcell effect, were observed for the first time in the solid phase.

In order to understand the use of quantum dots in this context, recall that, in the standard weak coupling regime where the emitter is coupled to a continuum of modes, the spontaneous emission rate is proportional to the

density of electromagnetic modes per unit volume. If an emitter is placed in an optical cavity, in resonance with one of its discrete modes  $M$ , and if this emitter is narrowly peaked from a spectral point of view, the mode  $M$  can be treated as a continuum. The density of modes associated with it is proportional to  $Q/V$ , where  $Q = \Delta\omega/\omega$  is the quality factor of the mode and  $V$  is the effective volume of the cavity. As the factor  $Q/V$  can take arbitrarily large values, this density of modes viewed by the emitter can be much greater than the density of modes in free space. More precisely, Purcell showed in 1947 that the spontaneous emission rate of the emitter in a cavity is enhanced or inhibited compared with its emission in free space by a factor

$$F_p = \frac{3}{4\pi^2} \frac{Q\lambda^3}{V},$$

where  $\lambda$  is the wavelength in the material. However, this is only valid if the emitter is perfectly coupled with the cavity mode.

The Purcell effect disappears if the emitter is spectrally shifted with respect to the strongly peaked maximum of the density of cavity modes, or if it is much broader spectrally than the cavity mode. QDs are particularly relevant in this context because their narrow spectral width makes it possible to use semiconductor cavities with high quality factor  $Q \sim 1\,000\text{--}10\,000$ . Enhancement factors for the spontaneous emission rate have been observed for InAs dots in the neighbourhood of 5 in micropillars [52] and of the order of 12 in microdisks [53].

Going beyond these first results, the association of quantum dots with very small volume semiconductor microcavities ( $V \sim \lambda^3$ ) is likely to set the scene for a wealth of new developments. The recent improvement in the quality of microdisks has opened the way to observation of the strong coupling regime for a single dot. In this regime, a photon emitted in the cavity mode is stored for long enough to be reabsorbed by the emitter before it can escape. Spontaneous emission then becomes a reversible phenomenon and the coupled atom/cavity system evolves in a deterministic way, even in the absence of an applied field. This system will offer very interesting prospects for quantum data processing [54].

The control of spontaneous emission in microcavities should also make it possible to reduce the threshold current in microlasers by several orders of magnitude. A threshold current of the order of the nanoampere is predicted at 300 K for an ensemble of quantum dots placed in an optical microcavity of ultimate size  $V \sim (\lambda/2)^3$ , fabricated in a photonic crystal, compared with at least  $30\ \mu\text{A}$  at the present time for the best surface-emitting lasers. In principle, it would even be possible to make a laser in which the active medium comprised a single quantum dot, and to lower the threshold current still further to a value of the order of  $10\ \text{pA}$  [55]. In this case, however, the emission from the quantum dot must be spectrally narrow enough to ensure good coupling of the cavity mode, and this will require low temperature operation ( $T < 100\ \text{K}$ ).

Although the technological obstacles are formidable, especially the electrical pumping of such nanosources, these estimates clearly illustrate the prospects for improving laser microsources that are opened up by applying the ideas of CQED to quantum dots. As far as applications are concerned, the miniaturisation of sources and reduction of their threshold current is a prerequisite for making dense optical ‘intra-chip’ interconnects in electronic circuits. The design of autonomous optical microsensors for biomedical or environmental applications is another goal often mentioned in this context.

## 16.3 Photonic Crystals and Microcavities

### 16.3.1 Introduction

Photonic crystals are periodic dielectric structures designed to modify the behaviour of photons in the same way as a crystalline material affects the properties of electrons. These structures should provide a way of significantly miniaturising components for optics and optoelectronics, with the aim of transporting, generating or handling a very large amount of data in a small space. More particularly, such objects may be used to produce a new generation of highly compact optical or optoelectronic components, such as wave guides, filters or microlasers.

These structures have periods of the order of the photon wavelength, i.e., a few hundred nanometers in the visible and near infrared ranges. It is now possible to make high quality photonic crystals on this length scale, thanks to the tremendous advances made in nanotechnology. In such a structure, photon propagation is impossible in a certain energy range known as the photonic band gap (PBG). The idea of a PBG was suggested in 1987 by E. Yablonovitch [56] and S. John [57]. One of the objectives was then to inhibit spontaneous emission by photon emitters located in the PBG of a photonic crystal.

Furthermore, a photonic crystal also has allowed bands in which propagation and conditional occupancy of photons is permitted. Their properties can then be closely controlled, i.e., wave vector, wavelength, group velocity, and lifetime. Photonic crystals are therefore able to enslave light in both time and space, at the scale of the photon wavelength.

After the pioneering work of the end of the 1980s, it soon became apparent that the very special properties of these structures could give rise to new, very compact components for use in optoelectronics. Many groups therefore developed theoretical and experimental approaches that would help to study and exploit these photonic crystals.

### 16.3.2 Periodic Structures

A photonic crystal is by definition a material comprising a crystal lattice of sites with dielectric constant  $\varepsilon_a$  immersed within a medium with dielectric



constant  $\varepsilon_b$  and with period of the same order as the wavelength to be controlled. We shall consider non-absorbent, linear, non-magnetic dielectric materials, i.e.,  $\mu_r = 1$ . In this case, Maxwell's equations reduce to

$$\begin{aligned}\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} && \text{Maxwell-Faraday,} \\ \nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t} && \text{Maxwell-Ampère,} \\ \nabla \cdot \mathbf{D} &= 0 && \text{Maxwell-Gauss,} \\ \nabla \cdot \mathbf{B} &= 0 && \text{Flux of } \mathbf{B} \text{ conservative,}\end{aligned}$$

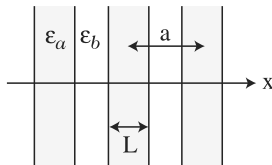
where  $\mathbf{D} = \varepsilon(\mathbf{r})\mathbf{E} = \varepsilon_r(\mathbf{r})\varepsilon_0\mathbf{E}$  and  $\mathbf{B} = \mu\mathbf{H} = \mu_0\mathbf{H}$ . We seek stationary solutions of these equation with a time dependence of the form  $e^{-i\omega t}$ . We concentrate on the magnetic induction field  $\mathbf{B}$ , which is transverse and thus simplifies the solution of the equations using the method known as the plane wave expansion [58]. The problem then reduces to solving the differential equation [59]

$$\nabla \times \left[ \frac{1}{\varepsilon_r(\mathbf{r})} \nabla \times \mathbf{B}(\mathbf{r}) \right] = \frac{\omega^2}{c^2} \mathbf{B}(\mathbf{r}),$$

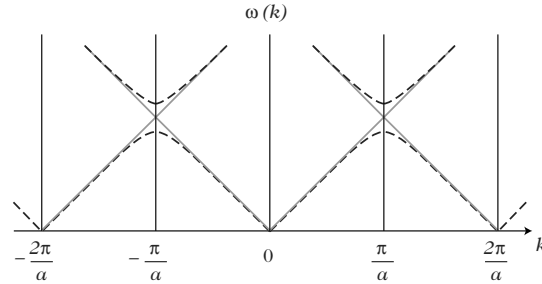
where  $c$ , which satisfies  $\varepsilon_0\mu_0c^2 = 1$ , is the speed of light. This equation plays an analogous role to the Schrödinger equation in the case of electronic structures. Since the photonic crystal is periodic, the solutions of this equation are Bloch functions  $\mathbf{B}_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}\mathbf{u}_{\mathbf{k}}(\mathbf{r})$ , where  $\mathbf{k}$  is the Bloch vector and  $\mathbf{u}_{\mathbf{k}}(\mathbf{r})$  has the periodicity of the photonic crystal [60,61]. The Bloch functions are periodic in the reciprocal lattice and we need only solve this equation in the first Brillouin zone. Taking into account the symmetry properties, we may in fact restrict to the high symmetry lines bounding the reduced zone (for more details, see [60,61]). We then obtain the structure of the photonic bands and the dispersion relation  $\omega(\mathbf{k})$ . To illustrate this, we shall begin by considering the simplest case of a periodic structure in just one dimension.

### 1D Photonic Crystals. Bragg Mirrors

Consider a 1D photonic crystal, also called a Bragg mirror, as illustrated in Fig. 16.19. If we examine the characteristics of the electromagnetic modes which have vector  $\mathbf{k} = k\hat{\mathbf{e}}_x$ , where  $\hat{\mathbf{e}}_x$  is a unit vector in the  $x$  direction, the equation to be solved is



**Fig. 16.19.** 1D photonic crystal made from parallel slabs with alternating dielectric constants  $\varepsilon_a$  and  $\varepsilon_b$  and period  $a$  in the  $x$  direction



**Fig. 16.20.** Photonic band gaps in a 1D structure. *Light grey curves:* Bands for a homogeneous structure, where the dispersion relation takes the form  $\omega k = cnk$ . *Dashed curves:* Bands of a 1D photonic crystal

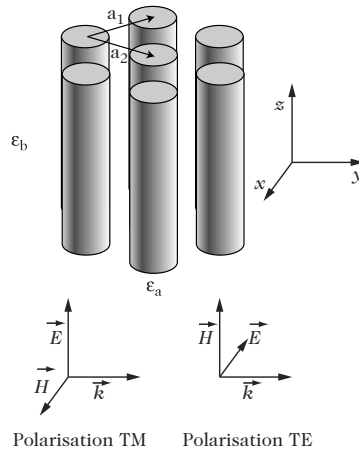
$$\frac{d}{dx} \left[ \eta(x) \frac{dB_k(x)}{dx} \right] = -\frac{\omega^2}{c^2} B_k(x),$$

where the function  $\eta(x)$  is the inverse of the relative dielectric constant  $\varepsilon_r(x)$  ( $\varepsilon_r$  being the square of the refractive index) and  $B_k(x)$  is a component  $\mathbf{B}_k(x)$  transverse to the  $x$  direction. One way of solving this equation consists in expanding in terms of a basis of plane waves (Fourier expansion), as is done in the electron case [60].

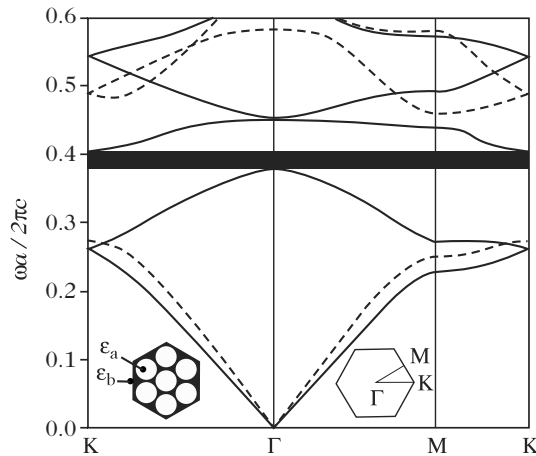
Figure 16.20 shows the appearance of photonic band gaps. The photonic bands are periodic in the reciprocal lattice, with period  $2\pi/a$ . The dielectric contrast removes the degeneracy of the bands at the edge of the Brillouin zone (at  $-\pi/a$  and  $\pi/a$ ). This removal of degeneracy creates an energy range in which the photons cannot propagate, called the photonic band gap (PBG). Since the removal of degeneracy is directly due to the contrast between  $\varepsilon_a$  and  $\varepsilon_b$ , materials with very different refractive index are required in order to open a broad PBG. It should also be noted that the dispersion relation is significantly modified near the PBG. In particular, the photonic band becomes ‘flat’, indicating that one can obtain very low group velocities  $v_g = d\omega/dk$ . This offers the prospect of very interesting applications for the storage and emission of light, as we shall soon see.

## 2D Photonic Crystals of Infinite Height

We now consider a structure that is periodic in two directions. We shall examine the ideal situation of 2D photonic crystals made from infinitely long cylinders or rods parallel to the  $z$  axis, as shown in Fig. 16.21. If we study the characteristics of electromagnetic waves propagating in the  $xy$  plane, the translation invariance in the  $z$  direction can be used to separate the field into two polarisations. In one of these, the electric field  $\mathbf{E}$  points in the  $z$  direction. This is the transverse magnetic (TM) polarisation. In the other, the magnetic field  $\mathbf{H}$  points in the  $z$  direction. This is the transverse electric (TE)



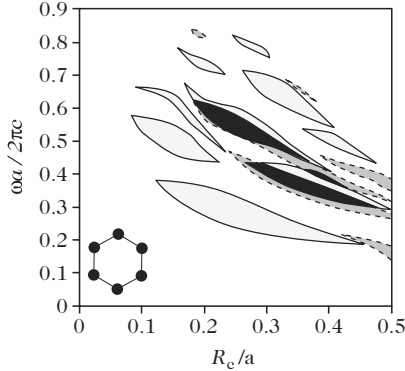
**Fig. 16.21.** Example of a 2D photonic crystal made from rods of dielectric constant  $\epsilon_a$ , infinite in the  $z$  direction and immersed in a medium of dielectric constant  $\epsilon_b$



**Fig. 16.22.** Photonic band structure of a photonic crystal made from cylinders of air ( $\epsilon_a = 1$ ) in a medium with dielectric constant  $\epsilon_b = 12.25$  (air filling factor 65%) for the TM polarisation (*continuous curves*) and the TE polarisation (*dashed curves*). The photonic band structure is given in reduced units ( $\omega a/2\pi c = a/\lambda$ ). The absolute PBG is indicated by a *black band*

polarisation. The band structures for these two polarisations are different and PBGs can be obtained for each of them.

For the photonic band structure illustrated in Fig. 16.22, there is a PBG for the TM polarisation between the first and second bands in the direction  $\Gamma M$ . However, there is no PBG in the  $\Gamma K$  direction (degeneracy is not removed at K for symmetry reasons). One speaks here of an incomplete PBG. In contrast, for



**Fig. 16.23.** Evolution of the PBG in a graphite-type arrangement of rods of radius  $R_c$  and dielectric constant  $\varepsilon_a = 13.6$  for the TM polarisation (*continuous curves*) and the TE polarisation (*discontinuous curves*). Absolute PBGs are shown in *black*

the TE polarisation, a broad PBG occurs between the first and second bands in all directions. This is a complete PBG, which totally forbids propagation in the  $xy$  direction. If one wishes to prevent propagation, whatever the polarisation, there must be an overlap between the PBGs of the two polarisations. One then speaks of an absolute PBG. It is important to note that the photonic band structure is given in reduced units  $\omega a/2\pi c = a/\lambda$ , where  $a$  is the period and  $\lambda$  is the wavelength in vacuum. This shows that there is a scaling law here. Hence, to obtain a PBG of given wavelength, it suffices to adjust the period of the photonic crystal. For the absolute PBG in Fig. 16.22, the period  $\alpha$  must be taken proportional to  $0.4\lambda$ .

Several parameters can be adjusted to obtain PBGs: the dielectric contrast, the arrangement of the rods in the crystal, and the shape of the rods. For example, Fig. 16.23 shows the dependence of the PBG on the radius of the rods in a structure with graphite-type arrangement [58].

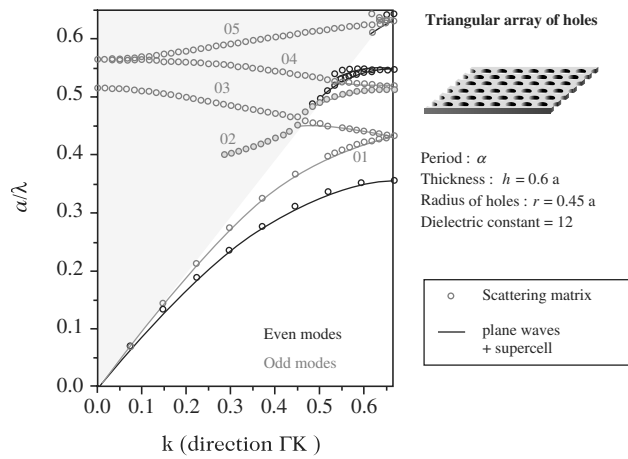
### Role of the Third Dimension

The commonest way of making a 2D photonic crystal begins with electron beam lithography, during which an array of holes is inscribed in a layer of organic material, followed by ion etching, which transfers the holes to the high index material. The etch depth is, of course, limited, and the result is often a long way from the case of infinitely long rods. One must then take into account the finite height of the rods in the third dimension. Indeed, the modes of the photonic crystal are then coupled with the continuum of radiative modes which can propagate in the air or the substrate surrounding the crystal.

Figure 16.24 shows a 2D photonic crystal of membrane type. In the photonic band structure as calculated by the scattering matrix method [62] and illustrated in Fig. 16.25, the continuum of radiative modes constitutes



**Fig. 16.24.** Illustration of a membrane-type 2D photonic crystal, consisting of holes in a dielectric surrounded by air both above and below



**Fig. 16.25.** Photonic band structure for a triangular array of circular holes in a suspended membrane of InP. The *shaded region* corresponds to the light cone [62]

a region called the light cone and the modes of the 2D photonic crystal within this continuum appear in the form of resonances. The light cone is bounded by a line called the light line, which corresponds to the dispersion relation  $\omega(\mathbf{k}) = ck$ .

For these quasi-3D photonic crystals, and as happens in standard wave guides, two categories of modes are distinguished:

- guided modes propagating within the guide zone and evanescent outside of it,
- radiative modes able to propagate in all the media.

Whereas the first (guided or bound modes), which are situated below the light line, will propagate without loss in the guide, the second (resonant modes), situated in the light cone, will undergo losses by evacuation into the surrounding

material [62]. The etch parameters, such as depth, geometry, or filling factor, can be adjusted to optimise the level of losses in the desired modes. For guided optics (fabrication of wave guides), one seeks to minimise these losses, whilst for antenna, LED, or laser design, one tries to maximise them, and this in certain favoured and predetermined directions.

### Different Materials. Weak and Strong Confinement

The required optical confinement properties, and in particular the expected modifications in the dispersion properties, place strict limits on the choice of dielectric media that could be used. Put briefly, the basic material must have both a high refractive index (typically around 3) and a controlled absorption coefficient. In addition, for certain components such as lasers or photodetectors, it is essential to use media with significant electro-optical effects.

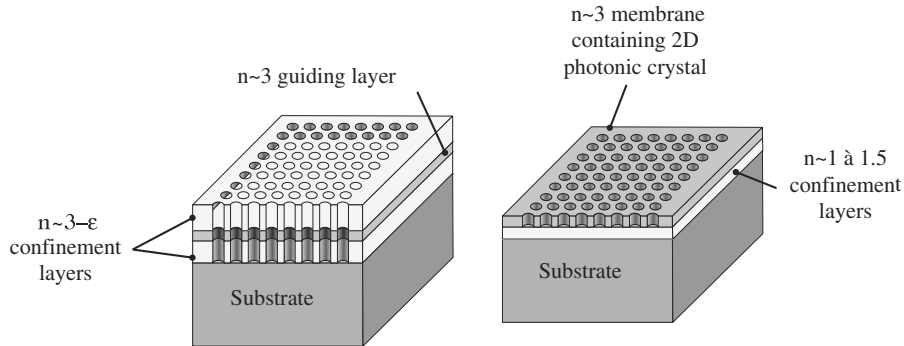
Given these prerequisites, the choice of semiconductor materials was soon made. Indeed, silicon and the compound III–V materials, especially GaAs and InP, have the required optical and electro-optical properties, and furthermore, they can be structured on the submicron scale, or even the nanoscale, using thin film technologies developed in microelectronics and optoelectronics. It is also possible to insert quantum wells, quantum dots, or impurities, whereby these materials can generate or absorb photons in a controlled way.

As already mentioned, it is then essential to confine the light within the thickness of the 2D photonic crystal. Since most proposed structures consist of air holes in a high index medium, two vertical confinement strategies have been put forward and abundantly discussed over the past few years.

The first recycles some of the achievements of integrated optics, using a wave guide with low vertical confinement, i.e., in which there is a low index contrast between a high index layer and the surrounding media (see Fig. 16.26). In such a configuration, most modes exist above the light line, intrinsically subject to losses, but the effective vertical radiation can be minimised. The III–V materials are perfectly suited to this approach, since one merely has to modify the composition of the alloy to go from the guiding layer, e.g., GaAs, to the confinement layer, e.g.,  $\text{Ga}_x\text{Al}_{1-x}\text{As}$ . This solution also provides a way of coupling the photonic crystals to standard optoelectronic components.

The second strategy makes use of the high index contrast between a semiconductor membrane and the surrounding medium (see Fig. 16.26). Vertical confinement is then drastic and modes can be exploited either above or below the light line. The first are called leaky modes and can be used to communicate between the photonic crystals and the surrounding space, e.g., through an optical fibre. The second are perfectly confined within the structure, and are therefore theoretically lossless. We may mention two types of technology which can achieve this rigorous vertical confinement:

- silicon-on-insulator (SOI) structures,



**Fig. 16.26.** Two available strategies for vertical confinement in 2D photonic crystals. *Left:* Weak confinement. *Right:* Strong confinement

- suspended InP membrane structures.

The second strategy is currently used by most groups working in the field of 2D photonic crystals. It has the greatest potential, both for producing novel physical effects and for developing new devices.

### 16.3.3 Structures Without Defects. Exploiting the Allowed Bands in Photonic Crystals

In 2D photonic crystals of finite height, bands can be engineered to adjust photon properties in a given volume: wavelength  $\lambda$ , wave vector  $\mathbf{k}$ , and lifetime  $\tau$ . Most of these parameters are determined by the band structure. Indeed, an optical mode whose wave vector has a component  $k_{XY}$  (related to the direction of  $\mathbf{k}$ , for a given wavelength) in the plane of the crystal will tend to couple with a radiated mode having the same wave vector (see Fig. 16.27). A photonic crystal component will communicate with the exterior through such a coupling. For example, in a laser, this is the direction that will characterise the output beam.

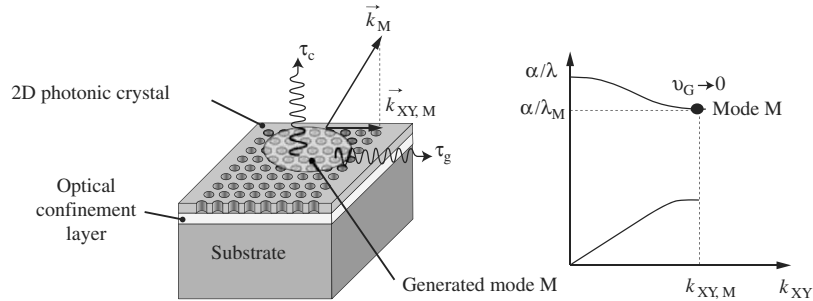
The lifetime for its part turns up in the quality factor of a resonator:  $Q = \omega\tau$ , where  $\omega$  is the angular frequency of the optical mode. This is related to the various possible loss mechanisms:

- in the plane of the photonic crystal ( $\tau_g$ ),
- losses by diffraction outside the plane, characterised by the time  $\tau_c$ .

In total, one finds

$$\frac{1}{\tau} = \frac{1}{\tau_c} + \frac{1}{\tau_g} .$$

Here,  $\tau_g$  depends on the group velocity of the mode, i.e., the slope of the dispersion curve under the chosen operating conditions. Only  $\tau_c$  can prove difficult to determine, sometimes requiring a specific electromagnetic model.



**Fig. 16.27.** Schematic views of a 2D photonic crystal (*left*) and its band structure (*right*). A mode M generated within the crystal can lose its energy in the crystal plane, thereby giving rise to guided modes, or outside the plane, whence radiated modes will be produced. The mode considered here is a resonant mode with lifetime extended by its very low group velocity. This mode can couple with radiative modes in the air in the direction of  $\mathbf{k}$

The most suitable operating conditions for a given application are thus found by analysing the band structure of the photonic crystal.

The choice of operating conditions obviously depends on what one intends to do with the device. In the case of a small laser emitter, the photons must be confined to a very small region for long enough to favour the local interaction of the optical mode with a gain material. In this context, it is primarily the quality factor  $Q$  that will be the determining parameter. Moreover, the direction of the output beam, as mentioned earlier, is determined by the wave vector of the resonant mode. We shall illustrate these two constraints later in the case of Bloch mode lasers.

In other applications, for example in the case of optical switching, one seeks to compel photons with some specific wavelength to choose a certain well determined direction of propagation. Hence, photonic crystals of superprism type make use of highly dispersive modes to distribute a polychromatic incident light beam over a large angular range. A different band structure will thus be required under such operating conditions.

If we focus on the example of laser emitters, the following two constraints are essential:

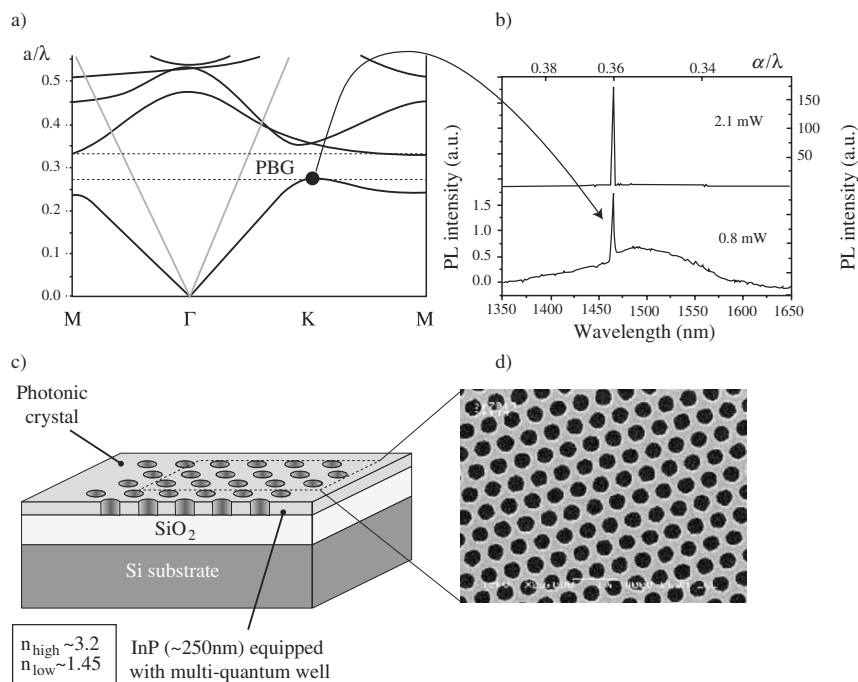
- adjustment of the quality factor,
- choice of output direction for the beam.

Furthermore, it is imperative to use a material with high optical gain. One must also carry out a spectral adjustment of the gain range and the confined mode of the photonic crystal. As an example, let us consider a photonic crystal laser whose design integrates these constraints. These are components which, like distributed feedback (DFB) lasers, exploit a periodic structure (period  $\lambda$ ) which generates a stationary wave capable of amplifying the optical mode–gain medium interaction. In terms of band structure, the stationary nature of this

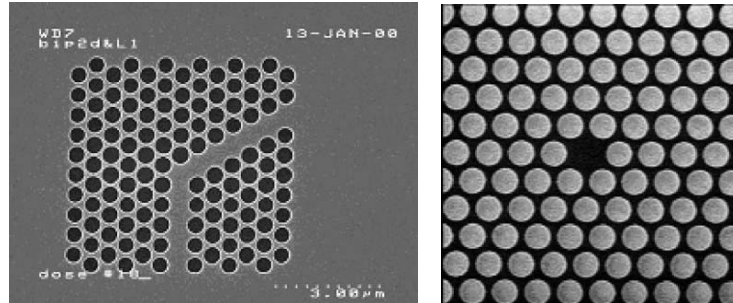


mode manifests itself through a band extremum for which the slope is zero. The resonance of this mode is characterised by its quality factor. However, in contrast to the standard DFB laser, the structures considered here are periodic in two dimensions and exhibit a very high degree of corrugation, i.e., strong refractive index contrast in the plane. This allows one to obtain a particularly clear extremum and a high level of localisation of the generated stationary mode.

Let us consider the extremum of the band structure in a triangular photonic crystal situated at the K point of the Brillouin zone, under the light line (see Fig. 16.28a). Theoretically, this operating point should give zero group velocity ( $\tau_g = \infty$ ) and zero diffraction losses ( $\tau_c = \infty$ ). A priori, confinement is thus perfect. Moreover, a stationary wave of this kind can couple with guided modes with high propagation constant. Such a structure is therefore well suited for making a monolithically integrated laser source with a photonic



**Fig. 16.28.** (a) Band structure of a 2D photonic crystal with PBG between the *dotted lines* and the light line shown in *grey*. The extremum at the critical point at K is indicated by a *black dot*. (b) Emission spectra. At the frequency corresponding to the K point, the emission peak is characteristic of a clear resonance which gives rise to laser emission above 1 mW. (c) Schematic and (d) SEM micrograph of a laser operating at the K point of the Brillouin zone



**Fig. 16.29.** *Left:* Guiding light through a photonic crystal. The region of the crystal in which holes have been omitted serves to guide the photons. *Right:* Confinement in a cavity of the photonic crystal. Photons emitted at the centre of the structure (region with no holes) are confined by the surrounding photonic crystal

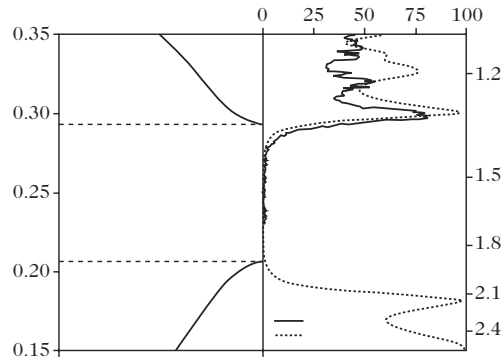
integrated circuit provided with wave guides. Figure 16.28 shows the experimental demonstration of this effect [63].

The structure is a photonic crystal fabricated on an InP membrane endowed with a multi-quantum well, capable of emitting photons around  $\lambda = 1.5\ \mu\text{m}$ . As a consequence, the lattice parameter is fixed at 530 nm. The membrane is mounted on silicon above a silicon host substrate. The experiment carried out here consists in generating photons by optical pumping. The photons are then stored in the resonant mode exactly where they are generated. In a certain sense, one thereby creates a self-positioned laser. Spectroscopic studies have revealed the resonant mode at  $1.46\ \mu\text{m}$  ( $a/\lambda = 0.36$ ), which coincides with the extremum predicted at K, up to experimental error. This mode has a quality factor of 800. In spontaneous emission, the mode is superposed upon the broad emission spectrum of the multi-quantum well. Beyond the threshold power of 1 mW, laser emission is achieved. The stimulated emission spectrum then exhibits a pure line corresponding to the resonant mode. The fact that one obtains laser emission with such structures implies that optical confinement can indeed be sharp in the three space directions with a simple 2D photonic crystal of finite height.

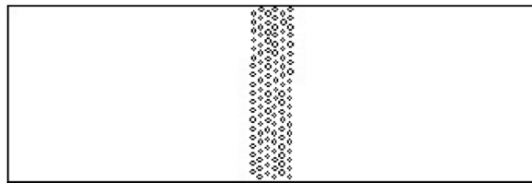
#### 16.3.4 Structures with Defects

We have just seen that it is possible to confine photons within a periodic structure without defects. However, there is another possibility if one uses the ability of photonic crystals to reflect and/or confine photons. One can then attempt to modify their propagation and guide them in a particular channel (see Fig. 16.29, left), or confine them in a restricted region of the semiconductor forming a cavity (see Fig. 16.29, right).

We shall give a more detailed illustration of these possibilities through three examples: a microcavity integrated into an etched wave guide, a hexagonal cavity, and a photonic crystal wave guide.



**Fig. 16.30.** *Left:* Calculated band diagram for the structure, showing that there is a band gap for reduced frequencies in the range 0.21–0.29. *Right:* Measured and calculated transmission spectra (*continuous* and *dotted curves*, respectively) for the photonic crystal. The wavelength range for which the light is not transmitted by the structure corresponds to the PBG



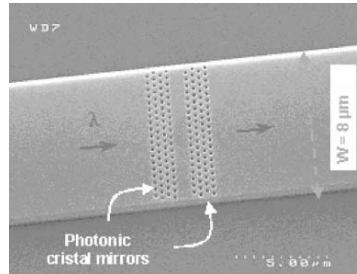
**Fig. 16.31.** Principle of the photonic crystal mirror

### Defects Serving as Cavities

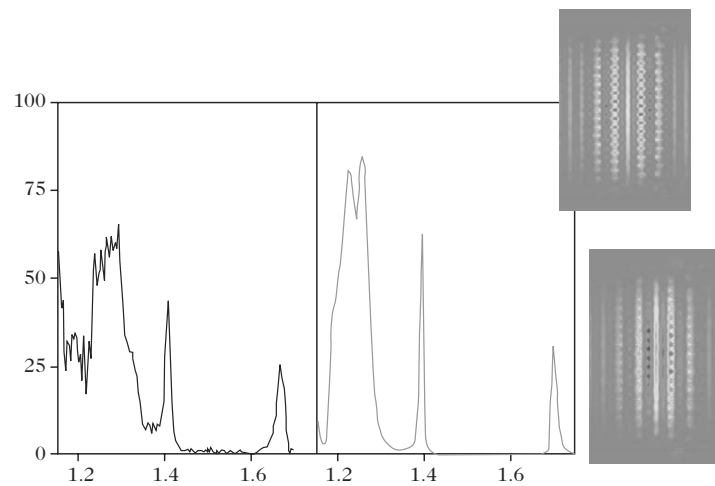
The first thing one can achieve here is to reflect the photons. To do this, one chooses a lattice, e.g., triangular, whose parameters, namely the period and filling factor, create a band gap at the desired wavelengths. An example band diagram for such a structure is shown in the left-hand part of Fig. 16.30. If this lattice is now fabricated in the middle of a wave guide, as shown schematically in Fig. 16.31, transmission through this mirror can be forbidden for photons propagating in the wave guide.

This is indeed what is effectively observed in the transmission spectrum shown on the right of Fig. 16.30. The light in the wave guide is no longer transmitted at wavelengths corresponding to the band gap. Encouraged by this success, we may now make a Fabry–Pérot resonator by fabricating two photonic crystal mirrors and placing them at a given interval to form a cavity (see Fig. 16.32). This creates a filtering function, since now only the resonant wavelengths of the Fabry–Pérot cavity will be transmitted by the structure, as can be seen from Fig. 16.33 [64].

One can see once again the band gap of the mirrors at which photons are no longer transmitted, but this time with transmission peaks corresponding to



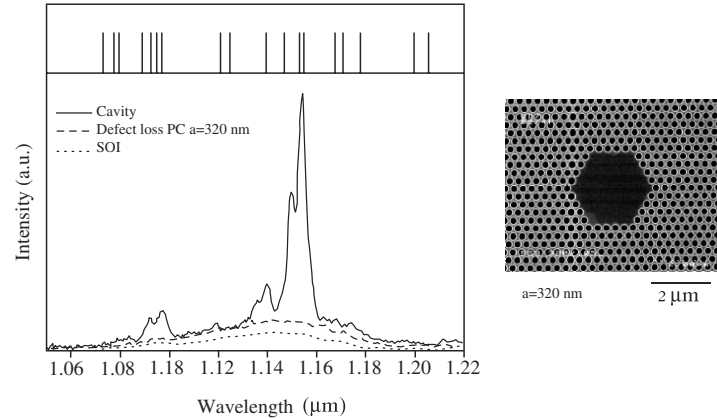
**Fig. 16.32.** Top view of an Si wave guide etched on an SOI substrate with width  $8\mu\text{m}$  and total length 10 mm. A photonic crystal microcavity has been fabricated within the wave guide by etching two photonic crystals (mirrors)



**Fig. 16.33.** Measured and calculated transmission spectra (black and grey curves, respectively) for the photonic crystal microcavity. Note the photonic band gap present in the case of a single mirror, together with the resonant modes of the microcavity which appear in this band. The electric field distributions for these two modes are also shown (right). They reveal the stationary nature of the resonant cavity modes [64]

the resonance wavelengths. The electrical component of the electromagnetic field in the structure (see Fig.16.33, right) confirms that these wavelengths do indeed resonate within the microcavity.

These same resonances can be obtained with different geometries, depending on what effect is required. For example, to privilege the interaction between photons and semiconductor, one generally attempts to keep the photons in as small a volume as possible and for as long as possible. One can then use the cavity as a resonator to hold the photons in a gain region and subsequently extract a great many of the photons generated in the semiconductor. These



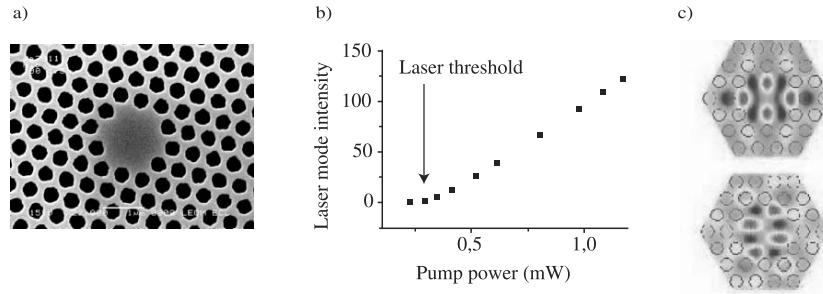
**Fig. 16.34.** *Top view* of a hexagonal cavity in a photonic crystal (*right*) and photoluminescence spectra (*left*) registered for an unpatterned part of the SOI (*dotted curve*), for the patterned part of the SOI constituting the photonic crystal (*dashed curve*), and for the cavity at the centre of the crystal (*continuous curve*). Note the significantly increased extraction of light in the case of the cavity, and also the spectral features due to resonant modes of the structure

are detected in the external medium, i.e., the air, in which the observer is located (see Fig. 16.34).

Moreover, with such cavities, it is possible to achieve laser emission, provided that the optical gain is sufficient to exceed optical losses from the cavity. At the present time, only heterostructures using III–V compounds of type GaAs and InP provide sufficient gain to obtain the laser effect. Figure 16.35a shows an InP microlaser transferred onto silicon equipped with a multi-quantum well which gives it a large optical gain around  $1.5\ \mu\text{m}$  [65]. It consists of a hexagonal microcavity of radius  $2\ \mu\text{m}$ . Optical losses are low enough ( $Q = 800$ ) to allow laser emission with an excitation threshold of only  $250\ \mu\text{W}$ , as can be seen from the gain curve in Fig. 16.35b. Figure 16.35c shows theoretical maps of the electromagnetic field of the degenerate mode from which laser emission is obtained.

These results show that the interaction between a photonic crystal cavity mode and a gain medium can already be exploited to make light emitters with sizes of the order of the photon wavelength. They open the way to the development of a new generation of light-emitting diodes and laser diodes. Apart from their small size, photonic crystals also allow one to sculpt the beam emitted by such a component, i.e., one can obtain very highly controlled directivity. In the case of lasers, it is also possible to reduce the emission threshold and hence also the energy consumption of the component.

Looking beyond these considerations, such structures are likely to reveal novel interaction effects between radiation and matter. In particular, one consequence of this ability to confine photons so strongly may be a radical



**Fig. 16.35.** (a) SEM micrograph of a photonic crystal cavity of diameter  $2\mu\text{m}$ . (b) Gain curve of the light emitter, from which the laser emission threshold can be determined. (c) Theoretical maps of the field of the degenerate mode associated with laser emission. *Dark lobes* represent regions where the field has high amplitude, i.e., where the light is confined [65]

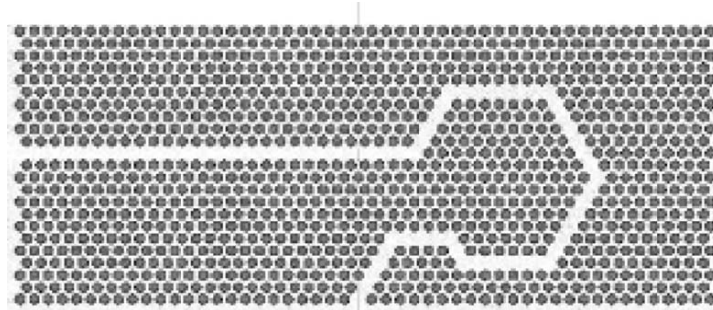
modification of the way photons are emitted, via the Purcell effect. Indeed, it is remarkable that using these cavities one can conserve such a high  $Q$  factor with a modal volume  $V$  that assumes a practically ultimate value, of the order of  $\lambda^3$ . This high value of  $Q/V$  should make it possible to inhibit or enhance the spontaneous photon emission rate by the Purcell effect, using suitable emitters such as semiconductor quantum dots and skillful engineering of defects in the photonic crystal. Finally, with such objects, it is also possible to enhance and explore nonlinear optical effects of  $\chi^{(2)}$  or  $\chi^{(3)}$  type.

Starting from active photonic crystals equipped with an optical gain medium, we may thus imagine the construction of original physical nanolaboratories. In the longer term, these should further widen the range of components that can be made from photonic crystals.

### Defects Serving as Wave Guides

We have just seen that photons can be reflected with the help of a photonic crystal. But it is also possible to guide them, as in a wave guide etched in a semiconductor. The difference is that, in a photonic crystal, the guiding effect will not be obtained by the difference of refractive index between the guide medium and its surroundings, but rather by a combination of reflection and diffraction of the photons at the interface of the guiding part, which is an unstructured part of the semiconductor, and the confining part, which is the photonic crystal itself.

We shall discuss here an example where the propagation of the electromagnetic field in an arbitrary structure has been calculated. As suggested by the representation in Fig. 16.36, it once seemed that photonic crystal wave guides would be able to include highly curved bends, which is generally a delicate matter in more conventional wave guides. This was one of the chief reasons for investigating such optical wave guides several years ago. However, despite



**Fig. 16.36.** *Top view* of a wave guide with several bends, made by omitting a row of holes. The theoretical distribution of the electromagnetic field in this structure shows that light is confined within the wave guide

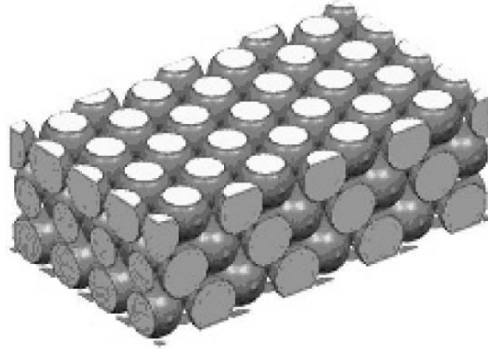
the efficient inhibition of radiation in the plane of the crystal, such bends are not without their drawbacks. In particular, they can cause a high level of light reflection. Moreover, propagation losses in such guides still remain higher than in conventional wave guides. So rather than simply reproducing optical functions that can be achieved by standard wave guides, these devices will only be able to realise their full potential by exploiting new functionalities. For example, as with 2D photonic crystals, the group velocity can be controlled. By slowing the photons down, coupling could be facilitated between guided and stationary modes, and this could be used to integrate filtering functions, for example.

### 16.3.5 Conclusion and Prospects

Photonic crystals were born from the simple idea that, by structuring matter in three dimensions, it would be possible to make a material that would behave in the same way with respect to photons as ordinary crystals do with respect to electrons. To obtain a PBG in all spatial directions, a 3D structure is thus required. But the fabrication of such structures using etching techniques still looks very difficult in the visible range of the spectrum.

A more economical possibility for fabricating such things consists in using self-organised systems. At the present time, the most interesting of these is the opal. This is a face-centered cubic arrangement of silica spheres, with variable diameter in the range 0.2–1  $\mu\text{m}$ , depending on the fabrication conditions, but a size dispersion below 5%. These spheres can be coated with a semiconductor to obtain more suitable refractive index contrasts. However, depending on the method of fabrication, opals exhibit various types of structural defect whose influence on optical properties remains to be studied.

At the same time, advances in micro- and nanotechnology have made it possible to fabricate two-dimensional structures. The first novel effects have been demonstrated, including the possibility of controlling the propagation



**Fig. 16.37.** Schematic view of an opal

speed of photons in a photonic crystal waveguide. Other effects seem to be within reach, such as the construction of a laser source without threshold. Further exciting prospects will certainly arise by applying the possibilities of photonic crystals in microfluidics and biology, insofar as there will always be a need to achieve combined optical, electronic and chemical functions within the same very small region of a semiconductor. One day we may hope to build genuine nanolaboratories, and perhaps even nanofactories!

## References

1. Tonks, L., Langmuir, I.: *Phys. Rev.* **33**, 195 (1929)
2. Raether, H.: *Surface Plasmons*, Springer Tracts in Modern Physics 111, Springer Verlag, Berlin (1988) p. 4
3. Nylander, C., Liedberg, B., Lind, T.: *Sensors & Actuators* **3**, 79 (1982); Liedberg, B., Nylander, C., Lundstrom, I.: *Biosensors and Bioelectronics*: **10**, i–ix (1995)
4. Homola, J., Yee, S.S., Gauglitz, G.: *Sensors & Actuators B* **54**, 3 (1999); Homola, J., Koudela, I., Yee, S.S.: *Sensors & Actuators B* **54**, 16 (1999)
5. Jorgenson, R.C., Yee, S.S.: *Sensors & Actuators B* **12**, 283 (1993)
6. Ebbesen, T.W., Lezec, H.J., Ghaemi, H.F., Thio, T., Wolff, P.A.: *Nature* **391**, 667 (1998); Ebbesen, T.: *Des photons passe-muraille*, *La Recherche* **329**, 50 (2000)
7. Barbara, A., Quémerais, P., Bustarret, E., Lopez-Rios, T.: *Optical transmission through subwavelength metallic gratings*, *Phys. Rev. B* **66**, 161403 (2002)
8. Hibbins, A.P., Sambles, J.R., Lawrence, C.R.: *Appl. Phys. Lett.* **81**, 4661 (2002); Lezec, H.J., Degiron, A., Devaux, E., Linke, R.A., Martin-Moreno, L., Garcia-Vidal, F.J., Ebbesen, T.W.: *Science* **297**, 820 (2002); Smolyaninov, I.I., Zayats, A.V., Gungor, A., Davis, C.C.: *Phys. Rev. Lett.* **88**, 187402 (2002)
9. Altwischer, E., Van Exter, M.P., Woerdman, J.P.: *Nature* **418**, 304 (2002)
10. Ghaemi, H.F., Thio, T., Grupp, D.E., Ebbesen, T.W., Lezec, H.J.: *Surface plasmons enhance optical transmission through subwavelength holes*, *Phys. Rev. B* **58**, 6779 (1998)



11. Ebbesen, T.W., Lezec, H.J., Ghaemi, H.F., Thio, T., Wolff, P.A.: Extraordinary optical transmission through sub-wavelength hole arrays, *Nature* **391**, 6668 (1998)
12. Kreibig, U., Vollmer, M.: *Optical Properties of Metal Clusters*, Springer, Berlin (1995)
13. Quinten, M.: *Appl. Phys. B* **73**, 317 (2001)
14. Jensen, T.R., Duval Malinsky, M., Haynes, C.L., Van Duyne, R.P.: *J. Phys. Chem.* **104**, 10549 (2000); Linden, S., Christ, A., Kuhl, J., Giessen, H.: *Appl. Phys. B* **73**, 311 (2001)
15. Nie, S., Emory, S.R.: *Science* **275**, 1102 (1997)
16. Kneipp, K., Kneipp, H., Iitzkan, I., Dasari, R.R., Feld, M.S.: *J. Phys. Cond. Matt.* **14**, R597 (2002)
17. Hillenbrand, R., Keilmann, F.: *Appl. Phys. B* **73**, 239 (2001)
18. Viets, C., Hill, W.: *J. Raman Spectr.* **31**, 625 (2000); Stokes, D.L., Vo-Dinh, T.: *Sensors & Actuators B* **69**, 28 (2000)
19. Krenn, J.R., Dereux, A., Weeber, J.C., Bourillot, E., Lacroute, Y., Goudonnet, J.P., Schider, G., Gotschy, W., Leitner, A., Aussenegg, F.R., Girard, C.: Squeezing the optical near-field zone by plasmon coupling of metallic nanoparticles, *Phys. Rev. Lett.* **82**, 2590 (1999); Krenn, J.R., Weeber, J.C., Dereux, A., Bourillot, E., Goudonnet, J.P., Schider, G., Gotschy, W., Leitner, A., Aussenegg, F.R., Girard, C.: *Phys. Rev. B* **60**, 5029 (1999); Krenn, J.R., Lamprecht, B., Ditzbacher, H., Schider, G., Salerno, M., Leitner, A., Aussenegg, F.R.: *Europhys. Lett.* **60**, 663 (2002)
20. Lamprecht, B., Leitner, A., Aussenegg, F.R.: *Appl. Phys. B* **68**, 419 (1999); Lamprecht, B., Krenn, J.R., Schider, G., Ditzbacher, H., Salerno, M., Felidj, N., Leitner, A., Aussenegg, F.R.: Surface plasmon propagation in microscale metal stripes, *Appl. Phys. Lett.* **79**, 51 (2001)
21. Dickson, R.M., Lyon, L.A.: *J. Phys. Chem. B* **104**, 6095 (2000)
22. Colombelli, R., Capasso, F., Straub, A., Gmachl, C., Blakey, M.I., Sergent, A.M., Sivco, D.L., Cho, A.Y., West, K.W., Pfeiffer, L.N.: *Physica E* **13**, 848 (2002); Tredicucci, A., Gmachl, C., Capasso, F., Hutchinson, A.L., Sivco, D.L., Cho, A.Y.: *Appl. Phys. Lett.* **76**, 2164 (2000)
23. Barnes, W.L., Dereux, A., Ebbesen, T.: *Nature* **424**, 824 (2003)
24. Podolskiy, V.A., Sarychev, A.K., Shalaev, V.M.: *Optics Express* **11**, 735 (2003)
25. Bergman, D.J., Stockman, M.I.: *Phys. Rev. Lett.* **90**, 027402 (2003)
26. Birotheau, L., Izraël, A., Marzin, J.Y., Azoulay, R., Thierry-Mieg, V., Ladan, F.R.: *Appl. Phys. Lett.* **61**, 3023 (1992)
27. Goldstein, L., Glas, F., Marzin, J.Y., Charasse, M.N., Leroux, G.: *Appl. Phys. Lett.* **47**, 1099 (1985)
28. Kirstaedter, N., et al.: *Electron. Lett.* **30**, 1416 (1994)
29. Huang, X., Stintz, A., Hains, C., Liu, G., Cheng, J., Malloy, K.: *IEEE Phot. Technol. Lett.* **12**, 227 (2000)
30. Kovsh, A.R., et al.: *Electron. Lett.* **35**, 1161 (1999)
31. Klopf, F., Reithmaier, J.P., Forchel, A.: *Appl. Phys. Lett.* **77**, 1419 (2000)
32. Schäfer, F., Reithmaier, J.P., Forchel, A.: *Appl. Phys. Lett.* **74**, 2915 (2000)
33. Matsuda, K., Ikeda, K., Saiki, T., Tsuchiya, H., Saito, H., Nishi, K.: *Phys. Rev. B* **63**, 121304 (2001)
34. Gérard, J.M., Cabrol, O., Sermage, B.: *Appl. Phys. Lett.* **68**, 1113 (1996)
35. Narukawa, Y., Kawakami, Y., Funato, M., Fujita, S., Fujita, S., Nakamura, S.: *Appl. Phys. Lett.* **70**, 981 (1997)

36. O'Donnell, K.P., Martin, R.W., Middleton, P.G.: Phys. Rev. Lett. **82**, 237 (1999)
37. Marzin, J.Y., Gérard, J.M., Izraël, A., Barrier, D., Bastard, G.: Phys. Rev. Lett. **73**, 716 (1994)
38. Kammerer, C., et al.: Phys. Rev. B. **66**, R041306 (2002)
39. Bayer, M., Forchel, A.: Phys. Rev. B **65**, 041308 (2002)
40. Brunner, K., et al.: Phys. Rev. Lett. **73**, 1138 (1994)
41. Landin, L., Miller, M., Pistol, M.E., Pryor, C.E., Samuelson, L.: Science **280**, 262 (1998)
42. Bennett, C.H., Brassard, G., Eckert, A.K.: Sci. Am. **267**, 50 (1992)
43. Brassard, G., Lütkenhaus, N., Mor, T., Sanders, B.C.: Phys. Rev. A **85**, 1330 (2000)
44. Kim, J., Benson, O., Yamamoto, Y.: Nature **397**, 500 (1999)
45. Moreau, E., Robert, I., Gérard, J.M., Abram, I., Manin, L., Thierry-Mieg, V.: Appl. Phys. Lett. **79**, 2865 (2001)
46. Gérard, J.M., Gayral, B.: J. Lightwave Technol. **17**, 2089 (1999)
47. Michler, P., Kiraz, A., Becher, C., Schoenfeld, W., Petroff, P.M., Zhang, L., Hu, E., Imamoglu, A.: Science **290**, 2282 (2000)
48. Santori, C., Fattal, D., Pelton, M., Solomon, G.S., Yamamoto, Y.: Phys. Rev. B **66**, 045308 (2002)
49. Moreau, E., Robert, I., Manin, L., Thierry-Mieg, V., Gérard, J.M., Abram, I.: Phys. Rev. Lett. **87**, 183601 (2001)
50. Purcell, E.M.: Phys. Rev. **69**, 681 (1946)
51. Haroche, S., Kleppner, D.: Physics Today **42**, 24 (1989)
52. Gérard, J.M., Sermage, B., Gayral, B., Costard, E., Thierry-Mieg, V.: Phys. Rev. Lett. **81**, 1110 (1998)
53. Gayral, B., Gérard, J.M., Sermage, B., Lemaitre, A., Dupuis, C.: Appl. Phys. Lett. **78**, 2828 (2001)
54. Imamoglu, A., Awschalom, D., Burkard, G., Di Vincenzo, D.P., Loss, D., Sherwin, M., Small, A.: Phys. Rev. Lett. **83**, 4204 (1999)
55. Pelton, M., Yamamoto, Y.: Phys. Rev. A **59**, 2418 (1999)
56. Yablonovitch, E.: Phys. Rev. Lett. **58**, 2059 (1987)
57. John, S.: Phys. Rev. Lett. **58**, 2486 (1987)
58. Cassagne, D.: *Matériaux à bandes interdites photoniques*, Ann. de Phys. **23** (4) (1998) p. 1
59. Joannopoulos, J.D., Meade R.D., Winn J.N.: *Photonic Crystals: Molding the Flow of Light*, Princeton University Press, Princeton, NJ (1995)
60. Kittel, C.: *Introduction to Solid State Physics*, 7th edn., Wiley, New York (1996)
61. Ashcroft, N.W., Mermin, N.D.: *Solid State Physics*, Holt, Rinehart and Winston, New York (1976)
62. Le Vassor d'Yerville, M.: Modélisation de cristaux photoniques bidimensionnels de hauteur finie, PhD Thesis, University of Montpellier II (2002)
63. Monat, C., et al.: Appl. Phys. Lett. **81**, 5102 (2002)
64. Zelsmann, M., et al.: Appl. Phys. Lett. **81** (13), 2340 (2002)
65. Monat, C., et al.: J. Quantum Electron. **39**, 419 (2003)

## Nanophotonics for Biology

J. Zyss and S. Brasselet

The term ‘nanophotonics’ has gradually been adopted to describe a new window for observing matter (see for example [1]). The prefix ‘nano’ refers to the nanometer length scale accessible through this new window ( $1\text{ nm} = 10^{-9}\text{ m}$ ), whilst the noun ‘photonics’ encompasses the whole set of ideas and applications of optics with a distinctive added value when compared with traditional optics, viz., a relevance to information processing and communications which opens new horizons for it. In the same way, according to a previous example whose success is still with us, electricity led to electronics with the advent of the transistor in the 1950s and 1960s. That innovation was based on the transport properties of the electron as an elementary vector for emitting, processing, and detecting information. In photonics, it is photons that replace the electrons of electronics and microelectronics as elementary vectors for the transport and processing of energy and information in the form of light. Building up on this new paradigm, roughly since the 1970s, photonics has been drawn forward by applications to telecommunications and more generally to information technology, a result of the fortunate cross-fertilisation between the semiconductor laser, born in the same period, and the technologies of microelectronics adapted to industrial developments in this new field of applications.

The founding and unifying principle of photonics, even more relevant today with the advent of high speed optical networks, is the guiding of light within optical microcircuits, itself resulting from physical concepts already proven on technologically less demanding millimeter or centimeter length scales, e.g., construction of radars or UHF circuits using metallic wave guides. Light can be confined on the scale of its own wavelength, i.e., the micron for the relevant wavelength range here (between the near UV and the near IR, via the visible part of the spectrum), by total reflection on an interface between two media with different refractive indexes. Over the past two decades, successive generations of silica optical fibres or optical wave guides of different kinds (doped glasses, semiconductors, polymers and hybrids) have marked out and illustrated the technological development of photonics, whose favoured scale

spans from a few  $\mu\text{m}$  to  $1/10$  or  $1/100\mu\text{m}$ . The lower bound corresponds, not to propagation phenomena as such, but rather to diffraction effects caused by defects or roughness, which must be controlled and reduced to a scale well below the wavelength scale.

Let us begin with an observation which is at first sight extremely restrictive, arising from the foundations of instrumental optics and carried forward by the precepts of quantum mechanics: the laws of diffraction forbid both the confinement and the propagation of light beams on any scale significantly shorter than the wavelength [2, 3]. In these conditions, adjoining the prefix ‘nano’ to the noun ‘photonics’ would appear to constitute a paradox, if not a contradiction, given the apparent impossibility of even producing, let alone propagating a wave of dimensions significantly smaller than  $\lambda$ . What we have here are sacred principles, like the Rayleigh criterion, a seemingly ineludible obstacle on the road to improving the resolving power of optical instruments, be they microscopes or telescopes. Incidentally, this same limit holds down the progress of silicon technology based on optical microlithography and slowly but surely drains the impetus of microelectronics as it moves down towards the fraction of a micron, condemning it to switch to other processes less propitious for industrialisation or the mass production of components at ever lower cost.

However, a deeper analysis reveals that there is nothing insurmountable about these confines, even if the basic principles remain fundamentally un-touchable. Indeed, a wave with at least one nanoscale dimension can be both generated and then detected in the far field by exploiting the phenomena of evanescent waves, first discovered in Newton’s prism experiments, but only really put to use over the past decade or so [4] with the advent of scanning near-field optical microscopy (SNOM). Other approaches have since appeared to consolidate this progress and diversify the panoply of nanoscale instrumentation, but these being based on the essential ingredients of biophotonics, it is perhaps the moment to stop and take a preliminary look.

The existence of tunneling optical waves and their application in such devices results from the well known phenomenon of reflection at an interface, which can be explained in the context of Maxwell’s theory of electromagnetism. Figure 17.1 shows evolutionary stages in the development from classical optics to one of the most symbolic achievements of nanophotonics. Indeed, if one writes down the full boundary conditions at a dielectric interface on which a wave is reflected in a total reflection configuration, i.e.,  $\theta > \theta_c$ , where  $\sin \theta_c = 1/n$ , reflection occurring in the half-space with refractive index  $n$ , the other half-space being occupied by air or a medium with lower refraction, with  $\theta$  the angle of incidence measured classically from the normal to the dioptric surface at the point of impact of the ray, one concludes as to the necessary existence of a wave with transverse profile and exponential decline on nanometric dimensions, from a few nanometers to submicron amplitudes, depending on the geometry of the experiment and the dielectric properties of the illuminated media.

This so-called evanescent wave is attenuated with exponentially decreasing amplitude in the direction perpendicular to the dioptric surface (in the half-space with lower refractive index, taken as air in this case, to simplify). It cannot therefore propagate, let alone be detected, without recourse to a trick pioneered by Newton himself, experimenter of genius. His idea was to bring a dielectric slab so close to the rear face of a prism, illuminated in a total internal reflection configuration, that it was almost touching. Today, we would say it was within tunneling distance. This new interface then serves to collect a significant part of the energy of the evanescent wave. The amplitude of the evanescent wave is captured by the propagating object, in this case the rear slab, typically half-way through its exponential decline, when its amplitude is of the same order of magnitude as the wave incident on the dioptric surface. It duly converts it to a propagating wave which can then be detected in the far field.

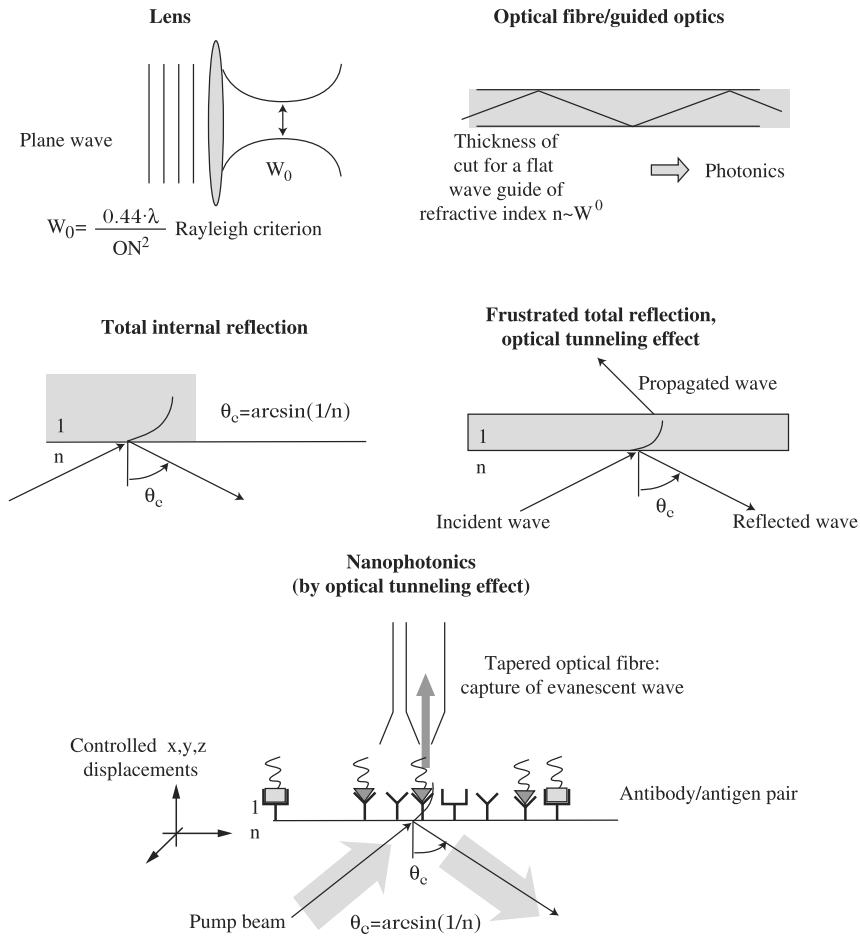
The result is a kind of frustrated total reflection on the first interface. But more importantly, a wave profile of nanoscale longitudinal dimensions (in the  $z$  direction, normal to the interface) has been rendered observable. The transverse dimensions remain those of a propagating wave profile generated and manipulated by means of conventional optics, e.g., by prisms, lenses, etc., and hence compelled to respect the constraints imposed by the laws of diffraction. We thus find that the problem of generating a nanoscale profile was partially resolved (in the  $z$  direction) by Newton before the end of the seventeenth century, albeit using other concepts and language. It remained of course to reduce the lateral length scales (in the transverse  $x$  and  $y$  directions), and this would take another three centuries or so with the advent of photonics, or more precisely, the invention of optical fibres. The relevant device is illustrated schematically at the bottom of Fig. 17.1.

It was indeed by tapering optical fibres until their transverse dimensions were reduced to at most  $100 \text{ \AA}$  [5] that lateral confinement could complement nanoscale longitudinal resolution with a transverse resolution of the same order. Using a feedback loop incorporating an electronically controlled piezoelectric device with  $xyz$  displacements, it became possible to achieve nanoscale imaging in the three spatial directions. Once the founding principles of nanophotonics had been laid and demonstrated via a generic tunneling microscope device, followed by other configurations, they were soon extended to the life sciences in the middle of the 1990s.

In this context, the related field of molecular photonics also played a determining role [6]. It treats of the light-matter interaction in an organic molecular medium, from the basic principles to the development of appropriate modern equipment and applications. The ideas and techniques in this area formed a productive association with those of nanophotonics, giving rise to nanobiophotonics (see for example [17]), which is the subject of this chapter.

The diagram at the bottom of Fig. 17.1 is a schematic view of a surface functionalised for biological purposes, with the aim of realising a nanobiophotoclip. The linguistic components of this ad hoc neologism will be

De l'optique à la Nano-Biophotonique



**Fig. 17.1.** Evolution of optics from its classical beginnings to the advent of nanophotonics as exemplified by the scanning tunneling microscope

explained below. For the solid-state physicist, there is no such thing as an ideal surface to materialise a perfectly flat dioptic surface. Indeed, the first nanophotonic devices that came with the generalisation of scanning tunneling microscopy provided a way of visualising asperities, defects, inclusions and unevenness of every kind, not to mention the various absorption phenomena which perturb and pollute the topography and composition of a wide range of sample surfaces. In this context, applications to biology are often based

on the intentional biochemical functionalisation of surfaces by deposition or binding of biological materials of all kinds (cells, proteins, neurones, etc.), with a view to observing or even processing them on a submicron scale that is quite inaccessible to classical microscopy, for the reasons discussed above.

The diagram at the bottom of Fig. 17.1 represents two types of antibody/antigen recognition system by squares (type I) and triangles (type II), corresponding to different diagnostic configurations. Indeed, using nanolithography, and in particular electron beam nanolithography as described in Chap. 1 or [8], or self-assembly techniques, as described in Chap. 2 or [9], surfaces can be functionalised according to a whole range of geometries and chemical compositions adapted to whatever purpose is at hand and on scales well below the microscale. These possibilities have led to new diagnostic technologies based on a combinatorial approach and on the specificity of antigen/antibody recognition mechanisms represented in the simplistic diagram by elementary triangle/square matching puzzles, sometimes referred to metaphorically as lock-and-key mechanisms. These will be discussed in more detail later in the chapter.

Now let us assume that, independently of any question of scale, the light-matter interaction leads to different fluorescence effects in the four cases which may occur in fundamental studies or diagnostic trials, namely the presence or absence of the complement of the antibody for structures I and II, the aim being to construct a nanoscale sensor or biophotoclip. Such differential fluorescence effects, obtained by incorporating distinctive fluorescent markers in the antigen (or the antibody, or both), should make it possible to unambiguously distinguish one of the four test configurations mentioned above. The different fluorescence situations can correspond on the one hand to the absence or the presence of fluorescence, and on the other, to distinct emission wavelengths, or again, emission wavelengths that are shifted with respect to their initial position in the absence of antigen-antibody coupling.

It remains to generate such fluorescence and detect it in imaging mode, with the nanometric spatial resolution required by the inherent ultracompactness of the diagnostic nanodevice whose operating principles are outlined here. It will be understood that the nanophotonic device depicted at the bottom of Fig. 17.1, which implements a tapered fibre, possibly metallised or emitting exciting light at the end [5], located above a substrate nanofunctionalised by biomolecules possessing specific fluorescence properties, is endowed with a set of characteristics which should be able to provide a joint answer to the question. Indeed, the exponential dependence of the evanescent profile in the  $z$  direction provides nanometric resolution in the longitudinal dimension, while lateral confinement in the capture by the optical fibre completes this resolution in the complementary transverse dimensions. The incident beam reflected from the dioptric surface is then a pump laser beam which excites the fluorescence of the markers attached to one or other of the molecular components to be detected.

More elementary configurations, but not less useful for the biologist, consist in depositing biological material such as a cell, a bacteria, or a tissue sample on the substrate, where ‘depositing’ means fixing and keeping it alive in the case of *ex vivo* observations or for the purposes of diagnosis.

Even though such devices, described in more detail in Sect. 17.3, do not constitute the only configuration so far developed in the context of nanobio-photonics, they do illustrate the key mechanisms, on both the fundamental and the practical level, which have allowed nanophotonics to flourish in the life sciences. They thus provide a good introduction to the rest of the chapter, as a specific illustration but also from a conceptual standpoint.

By describing the underlying principles of such a device, we have thus been able to indicate several of the more significant historical landmarks, and at the same time, to situate nanophotonics within a more complex ensemble of technologies and phenomena covering a range of length scales. Concerning the historical aspects, it is doubtless worth recalling that, although advances occur in the form of discontinuous leaps and bounds, generally at some lucky crossroads between a creative personal vision and contemporary technical knowhow capable of supplying an appropriate solution, such progress can never take place *ex nihilo*, but is always based on previous advances. This has been the case with nanobio-photonics, which came into being as a result of earlier, often largely disconnected work in photonics (fibre optics, lasers, wave guides, etc.) and molecular physics (light–matter interaction, absorption and fluorescence spectroscopy, functional molecules for optics), as illustrated by the genesis of the SNOM device in the figure. If one accepts that biology and related areas were able to grow into a genuine scientific discipline mainly thanks to the successful development of microscopy at the beginning of the seventeenth century, opening up a new observational window on the scale of cellular entities such as bacteria, it is not difficult on the same basis to assess the implications of the quantitative and qualitative leap in spatial resolution across several orders of magnitude that is the trademark of nanophotonics. For this step has indeed provided access to the molecular and subcellular genetic intimacy of biological mechanisms underlying the various pathologies. This progress in spatial resolution has been complemented in the time domain by the advent of ultrashort pulse lasers (see for example [10]). This has driven time resolutions down to the femtosecond level ( $1 \text{ fs} = 10^{-15} \text{ s}$ ), which corresponds to a fraction of an optical cycle in the visible.

Even though this is still a field in its very early stages, offering unpredictable prospects that will depend on the future advances of photonics in the wider sense, current benefits in the area of fundamental science (sequencing, proteomics, virology, neurology, etc.) exemplified by the first steps into the associated areas of prevention, diagnosis, and intervention, fully justify the convergent involvement of a growing number of research groups. These include physicists, chemists, and biologists working on projects in which the cross-disciplinary aspect belongs to an everyday reality that is already well established, going well beyond any simple question of fashion. (In this context,



it is worth remembering that the interface between physics and the life sciences, often presented as a new field of scientific activity, is in fact the result of a long tradition going back to pioneers like Volta, Galvani, or Helmholtz. However, this takes nothing away from the constantly renewed vitality at the frontiers of these fields.)

In the context of a discussion which appears to be exclusively focused on the nanoscale, it is important to emphasise the multiscale aspect intrinsic to nanobiophotonic studies. Although access to the nanoscale is indeed the new frontier to be explored in this area, there would be no way of doing so, either on the conceptual level or experimentally, without reference to complementary length scales, in particular the microscale, as has already been illustrated through the description of near-field observational methods. One must therefore avoid any reductional approach focusing exclusively on the nanoscale, without reference to the other levels making up the whole. For only such a multiscale approach can account for the highly entangled nature of the multiscale phenomena at work in both biological and physical media, or indeed, make it possible to implement physical effects and equipment involving all these scales in a joint manner.

In this context, Tables 17.1 and 17.2 distinguish and illustrate the four relevant length scales which together underlie the multiscale complexity of the phenomena under investigation and the corresponding techniques. They do so through the various supporting fields that are essential to photonics when applied to living beings: physics, chemistry, biochemistry, and of course, biology itself. These length scales are:

- the subnanoscale – the basic atomic and molecular scale of the order of a few Å,
- the nanoscale – from the nanometer to a few tens of nanometers,
- the mesoscale – of the order of a hundred nanometers,
- the microscale – length scale of typical electromagnetic wavelengths, going from the near UV, at several hundred nanometers, to the near or mid-infrared, sometimes referred to by the corresponding THz frequency range, i.e.,  $10^{12}$  Hz, which is situated on the length scale of a few tens of microns.

This generally accepted classification corresponds, as can be seen from Tables 17.1 and 17.2, to physical effects or material systems that can be clearly identified and exhibit basic features specific to each of the four levels. However, it is hard to draw a clear line between these domains, which are mainly intended therefore to establish orders of magnitude for reference purposes rather than strict boundaries. This is the case, for example, for the boundary between the nanoscale and the mesoscale, which we have chosen to distinguish here, even though they clearly overlap at the upper limit of the nanoscale, whilst each domain involves relatively distinct concepts and phenomena.

The physical, chemical and biological structures and phenomena related in some way to photonics are thereby listed in connection to the relevant length scales in the two tables. Table 17.1 considers the underlying concepts

**Table 17.1.** Length scales and their relationship with relevant branches of science. Underlying concepts and structures

	Atomic and molecular ( $\text{\AA}$ to a few nm)	Nanoscale (a few nm)	Mesoscale (a few 100 nm)	Microscale (up to a few tens of $\mu\text{m}$ )
Chemistry	Functional molecules for photonics	Supramolecular chemistry, interactions, delocalisation,	intermolecular nanoparticles	Micropatterned materials
Biochemistry	Elementary molecular building blocks of life: bases and amino acids	Specific molecular recognition	Structuring feedback	Self-organisation
Physics:				
Space and matter	Molecular orbitals X and synchrotron	Excitons and quantum transport phenomena,	confinement, nanoparticles	Thin layers, integrated optics
Radiation and spectrum	New physics of single molecules	Evanescence waves, local field, UV		Electromagnetic confinement visible $\rightarrow$ IR and UHF
Biology	Biochemistry	Cell compartments, sequencing, proteomics, virology		Cells and intercellular signalling, tissue

**Table 17.2.** Length scales and their relationship with relevant branches of science. Examples. NLO = nonlinear optics, GFP = green fluorescent protein, RFP = red fluorescent protein

	Atomic and molecular (Å to a few nm)	Nanoscale (a few nm to 100 nm)	Mesoscale (100 nm to a few 100 nm)	Microscale (up to a few tens of μm)
Chemistry	Molecular engineering for photonics: functional atoms and molecules (laser, NLO, luminescence, multifunctionality)	Intermolecular interactions (Van der Waals, Forster, Frankel). Clusters, highly delocalised molecular orbitals, supramolecular chemistry (cryptands, etc.)	As for nanoscale. Dendrites, electron delocalisation, mesophases	Polymers, thin layers, hybrid materials, photochemistry, self-organisation
Biochemistry	Molecular engineering for biophotonics: nonlinear luminescent markers, metallic ions and lanthanides	Photonic modification of genetic material: GFP, RFP, etc. Molecular recognition	pNAs (artificial analogs of oligonucleotides), micelles	Biopolymers (actin, etc.), artificial membranes
Physics: matter and radiation	X and synchrotron, atomic physics, elementary clusters	Extreme UV, molecular exciton (Frankel), quantum confinement, nanoparticles, local field effect, evanescent wave	UV, delocalised excitons (Wannier), polaritons, assemblies of nanoparticles	Visible, infrared, electromagnetic confinement ( $\sim \lambda$ ), integrated optics and optical fibres, photonic crystals
Biology	Amino acids, bases and pairs of bases (biochemistry)	Oligonucleotides, genetic material, ribosomes, etc.	Subcellular compartments, DNA, proteins, enzymes, membranes, viruses	From nucleus to cell, bacteria, cytoskeleton

and structures, whilst Table 17.2 collects several more specific examples. The reader will thus be able to situate nanobiophotonics at the meeting point of a much broader set of multiscale disciplines and issues connected with matter – inert or living – and radiation – external (exogenous) or emitted by the matter under investigation (endogenous).

One of the principal issues in chemistry is the design and construction of functional molecular structures, targeted for use in nanophotonics, making rational use of an unlimited supply of existing or synthesisable molecules from individual chromophores to supramolecular structures and polymers, and taking advantage from the double assistance of both experiment and theory. Each of these families of molecules is in itself a considerable resource. Polymers can serve as a material support for chromophores, bestowing upon them a structural added value that is extremely useful for the cohesion and behaviour of usable materials. On the other hand, they may themselves be endowed with emission properties, as attested by the present upsurge in organic light-emitting diodes (based in part upon conjugated structures suitable for charge transport and radiative recombination), which have now reached the mass production stage.

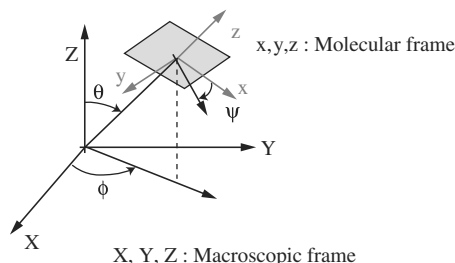
This vast area, which will be discussed in more detail in Sect. 17.2, comes under the heading of molecular engineering. It was first developed in the 1980s and 1990s in the context of information technology (confinement of light and integrated optics, nonlinear optics and electro-optics, luminescence, and lasers, the general principles of which are described in Sect. 17.1). Since the middle of the last decade, it made a further leap forward in the context of nanophotonics and in particular, its applications to biology. It is in this way that two-photon fluorescence, or the generation of second and third harmonics, initially developed as physical concepts for application to telecommunications [11], can now be found in only slightly altered form to observe and monitor the elementary mechanisms of life.

At the origin of this new trend lies a crucial event which occurred at the beginning of the 1990s and which in itself constitutes a genuine conceptual and methodological revolution with regard to the prospects for photonics, opening up a quite unsuspected field of application in the life sciences. Indeed, the traditional way of designing materials at the molecular level proceeds according to the so-called bottom-up approach, moving from the microscopic to the macroscopic, i.e., first designing functional molecules and then organising them on a larger scale in such a way as to reveal at the supramolecular level the initial property buried on the molecular level. (This organisation may be spontaneous in the case of crystal structures or mesophases, or it may be assisted in the case of materials structured mechanically, or under electrical or optical fields.) A case study with far-reaching applications is the breaking of centrosymmetry, which must be guaranteed both at the microscopic level of the molecule and the macroscopic level of the material. This is discussed in more detail in Sect. 17.1.

At the final level of the macroscopic structure, one may be dealing with weakly or strongly oriented statistical ensembles such as polymers or micelles, or in contrast, with perfectly regularly patterned arrays such as one finds in crystals. In the first case, that of rather poorly condensed statistical constructions (or soft matter, a rather unfortunate term, since these materials may turn out to be rather robust, or even more robust than the mineral, precisely by virtue of their flexibility and adaptational features, especially in living matter), the main conceptual tool was statistical physics, and until recently, experimental capabilities only gave access to ensembles involving a very large number of entities (at least of the order of a million for the small aggregates that can be observed using classical microscopy). These observations therefore led to statistical averages, such as the statistical moments of physical observables. The simplest and most common example is the dipole, which is the spatial moment of order 1 of the electronic and nuclear position variables. Higher order moments are referred to as multipoles to be discussed extensively in the rest of the chapter.

Of course, there is no doubt that the combination of the powerful conceptual edifice of statistical physics and the continual refinement of physical techniques, both with regard to preparation and observation, has led to spectacular progress for over a century now. However, our understanding of the interaction between light and matter has always come up against an insurmountable barrier with regard to the intrinsic behaviour under irradiation of the basic building blocks, i.e., individual atomic or molecular objects, which were thought to escape any direct observation. Hence, even if in the physical sciences we stubbornly maintain a Platonic stance in which we are condemned to see at best the reflected image of an ever-elusive reality (the search for a primitive object in its essence may be a suggestive illusion, but it is nevertheless an illusion), this limitation is relative and the quality of images can always be further improved. There always exist more sophisticated 'cave walls' on which the projected image may appear more clearly and better resolved. And indeed, the image itself is no doubt preferable to an image of an image or the idea of an image.

In this respect, a corner of the veil concealing the reality of the single molecule was lifted at the end of the 1980s by M. Orrit, then at Bordeaux (France), and W.E. Moerner, then at IBM San Jose (USA) (see [12] for an up-to-date and objective historical account). This breakthrough built upon previous work in fundamental physics and its applications (to be exact, spectral hole burning in its early quest to realise very high density optical memories), research that had opened up a new area in the spectroscopy of single objects. These studies began with the absorption of light by cold, isolated molecules, an extraordinary feat in itself but a shortlived success, and then eventually led to fluorescence studies, which lie at the root of the tremendous progress since achieved in single-molecule physics and biology. The obstacle imposed by low temperatures, which would have been as fatal for biological applications as it was for optical memories, was eventually overcome.



**Fig. 17.2.** Euler angles for description of molecular orientations

A large part of Sect. 17.2 will be devoted to this truly revolutionary achievement. Indeed, it breaks completely with the bottom-up approach mentioned earlier (which remains very much on the cards for the realisation of materials and devices). It opens a new window on single objects whose unprecedented applications and conceptual range are continually expanding. The first state-of-the-art experimental setups used cryogenic equipment, inherited from research the decade before into spectral hole burning, but this requirement soon gave way in the mid-1990s to one- or two-photon confocal microscopy for the main part, whereby liquid or condensed phase materials, including biological samples, became accessible with operation at room temperature.

On a more conceptual level, it became possible to test, confirm or invalidate previous work based on the more global models of statistical physics through the ability to sample statistical distributions on the level of their individual molecular constituents, which could only be probed in a collective manner prior to that [13]. Physicists thus began to study materials that form the very basis for the natural structures and phenomena related to life and which were previously out of bounds with respect to their field of investigation, i.e., inhomogeneous, poorly condensed, poorly ordered and fluctuating media.

It thus transpires that, although nanobiophotonics gains from its situation within a more extensive multiscale landscape, it is no less the crucial link for a great many present and future developments and represents a new frontier to be conquered at the confines of physics and the life sciences. The aim of the three main sections of this chapter will therefore be to examine these developments by introducing the underlying principles and presenting a choice of particularly symbolic experimental achievements to illustrate the state of the art. The reader will thus be prepared, if he or she should wish it, to confront the many challenges that will incite the physical and life sciences to join forces, using the tools described here, or better still, using those that remain to be developed, and to which he or she may contribute.

## 17.1 Emission and Absorption of Light by Molecular Systems

Biophotonics research makes use of interaction processes between light and matter which result in absorption and/or emission of radiation. In the most

general case, the matter interacting here is composed of a statistical ensemble of molecules characterised by an orientational distribution  $f(\Omega)$ . This function represents the probability distribution of molecular orientations in a unit volume and may be spatially dependent in the case of inhomogeneous content such as living tissue or cells. One would then write  $f(\mathbf{r}, \Omega)$ , but we discard this notation for the time being, for the sake of simplicity. The random variable  $\Omega$  is the solid angle characterising the orientation of a random molecular frame of reference with respect to fixed laboratory axes. This variable  $\Omega = (\theta, \varphi, \psi)$  can be defined by the Euler angles which relate the relative orientations of the macroscopic and molecular frames as shown in Fig. 17.2. Expressions for light fluxes are then given by orientational averages found by integrating the relevant physical observables weighted by the distribution  $f(\Omega)$  with respect to the random variable  $\Omega$ . One can also deal with isolated single entities, in which case no integration is required, but the orientation of the target molecule must then be known.

One of the main reasons for implementing light-matter interactions involving one or several photons is precisely to determine the orientation of individual molecular systems within environments of different kinds and on different scales, either by determining the a priori unknown distribution  $f(\Omega)$ , or by determining as accurately as possible the orientation of a single functional object, such as a light emitter or a nanoparticle attached to an oligonucleotide (see Sect. 17.2). The partial redundancy existing between the various processes described below lends itself well to such an approach. A coordinated or even simultaneous implementation can, in the case of a statistical ensemble of molecules, establish the coefficients in the Wigner expansion of  $f(\Omega)$  to a good approximation. This is the angular analog of the Fourier expansion for spatiotemporal distributions, and it leads to an expression of the form

$$f(\Omega) = \sum_{m', m, J} f_{m', m}^J D_{m', m}^J(\Omega), \quad (17.1)$$

where the basic distributions  $D_{m', m}^J(\Omega)$ , known as the Wigner functions, are tabulated and generalise the spherical harmonics with two angular variables (spherical coordinates) to the Euler angles (three angular variables). The indices  $m$ ,  $m'$  and  $J$  are the orders associated with the angles  $(\theta, \varphi, \psi)$ , where  $-J \leq m, m' \leq J$ . A combination of the processes mentioned below can be used to determine the coefficients  $f_{m', m}^J$  up to order 6. This Wigner expansion can be viewed as the spatial (angular) analog of the Fourier expansion in the time domain, the Wigner angular functions playing the same role as the functions  $\sin(\omega t)$  and  $\cos(\omega t)$ .

For all the processes described here, one assumes that the interaction processes with light have a purely intramolecular origin. This amounts to neglecting at this stage the consequences of collective effects resulting from intermolecular interactions. The latter are by nature weaker than the former. However, exciton effects (and their mixed extensions of polariton type, etc.) may play a significant role in the context of a strong interaction (in the Rabi

sense), as happens for example in optical microresonators, a subject beyond the scope of the present discussion. Another important domain in which intermolecular interactions are dominant will be discussed later in this chapter under the name of FRET (Förster resonance energy transfer). With these approximations, which are still valid in many situations of interest in biophotonics, one may consider that the only relevant physical observable for the following set of processes is the time and space dependent molecular dipole induced by excitation due to the electromagnetic field irradiating the medium. This observable corresponds to the instantaneous separation between the centre of gravity of the electrons in the molecule, which reduces in practice to the most polarisable  $\pi$  electrons, and the centre of gravity of the nuclei, which is treated as fixed in the context of the Born–Oppenheimer and Franck–Condon approximations, valid for short intervals in the aftermath of ultrashort pulses. When there is no external field, this separation results from intramolecular (electron–nucleus and electron–electron) interactions which lie at the origin of the permanent dipole moment  $\mu_0$ . Under irradiation conditions corresponding to normal fluence values, i.e., of the order of ten to a few hundred  $\text{MW cm}^{-2}$  away from resonance, delivered by a laser field  $\xi^\omega$  at frequency  $\omega$ , a perturbation expansion of the following kind is justified:

$$\begin{aligned} \mu = \mu_0 + \alpha : \xi^\omega + \beta_{\text{SHG}} : \xi^\omega \otimes \xi^\omega + \gamma_{\text{THG}} : \xi^\omega \otimes \xi^\omega \otimes \xi^\omega \\ + \gamma_{\text{NLI}} : \xi^\omega \otimes \xi^\omega \otimes (\xi^\omega)^* + \dots \end{aligned}$$

The subscripts NLI, SHG and THG will be explained shortly. Alternatively, one can project along the  $i$  axis:

$$\mu_i = \mu_{0i} + \sum_j \alpha_{ij} \xi_j^\omega + \sum_j \beta_{\text{SHG}_{ijk}} \xi_j^\omega \xi_k^\omega + \sum_j \gamma_{\text{THG}_{ijkl}} \xi_j^\omega \xi_k^\omega \xi_l^\omega + \dots \quad (17.2)$$

The tensor notations  $:$  and  $\otimes$  denote the partially contracted product and the tensor product, respectively. The standard notation  $:$  indicates the operation known as partial contraction between the linear polarisation tensor  $\alpha$  or nonlinear polarisation tensors  $\beta, \gamma$ , etc., and the corresponding field tensor, leading to a vector quantity. For example,  $\beta : \xi^\omega \otimes \xi^\omega$  is a vector with  $i$  components  $\sum_{jk} \beta_{ijk} \xi_j \xi_k$ , just as  $\gamma : \xi^\omega \otimes \xi^\omega \otimes \xi^\omega$  corresponds to a vector with  $i$  components  $\sum_{jkl} \gamma_{ijkl} \xi_j \xi_k \xi_l$ , and so on.

In this expansion, it is important to understand the tensorial nature of the polarisability coefficients  $\alpha, \beta, \gamma$  (Cartesian tensors of ranks 2, 3, and 4, respectively, given in a microscopic frame by their Cartesian components  $\alpha_{ij}, \beta_{ijk}$ , and  $\gamma_{ijkl}$ ), and also the products of the electric fields whose Cartesian components are given by  $(\xi^\omega \otimes \xi^\omega)_{ij} = \xi_i^\omega \xi_j^\omega$ , etc. These tensorial properties relative to the matter and the field, respectively, are of great importance in the highly anisotropic context of biochemical organic molecules. The polarisation states of the fields interacting with such entities will be able to probe



to high orders associated with nonlinear effects, i.e., not only to rank two tensors corresponding to the linear anisotropy, which would be a severe limitation given the complexity in the shapes of molecules proper to life, since it amounts to assimilating them in a rather gross manner to ellipsoids (with all due respect to approximation as a basic methodology in the physical sciences, albeit reaching its limits in the complex situations often encountered in the life sciences).

To simplify, we have assumed here that the radiation comes from a monochromatic laser, or a laser whose spectral width  $\delta\omega$  in angular frequency is small compared with the band widths of the molecular energy levels involved in the interaction. The electric field  $\xi^\omega$  considered here is not the external field irradiating the medium, but a so-called local field related to the latter by correction factors assumed to be incorporated in the appropriate way into the coefficients  $\alpha$ ,  $\beta$ , and  $\gamma$ , and whose weight increases with the order of the term considered in the dipole expansion. These correction factors are given to a good level of approximation in most situations by the Lorentz–Lorenz model. This treats each molecule as sitting within a ‘gedank’ cavity carved out from a linearly or nonlinearly polarisable external continuum and containing a discrete and countable set of molecules. The expansion coefficients then take the values

$$\alpha = \alpha^\circ F^\omega, \quad \beta_{\text{SHG}} = \beta_{\text{SHG}}^\circ F^{2\omega} (F^\omega)^2, \quad \gamma_{\text{THG}} = \gamma_{\text{THG}}^\circ F^{3\omega} (F^\omega)^3, \quad (17.3)$$

where

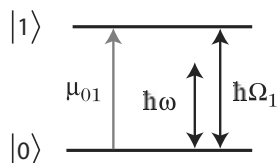
$$F^{m\omega} = \frac{(n^{m\omega})^2 + 2}{3}, \quad m = 1, 2, 3, \dots,$$

and the suffix  $\circ$  refers to a molecular entity assumed to be ideally isolated. (In the following, we shall omit the suffix  $\circ$  to simplify the notation.) Note that the dipole expansion given above is expressed in a Fourier space and results from prior considerations in the direct (temporal) space [14]. The dipole–field relationship then appears in the form of multiple spatiotemporal convolutions based on the requirements of causality, time invariance, and locality of the response of the molecular system.

In (17.2), the complex Fourier components of the field can be decomposed into amplitude and phase according to

$$\xi^{m\omega} = \mathbf{E}^{m\omega} e^{jm\omega t}. \quad (17.4)$$

This reveals, in the reduced dipole spectrum, nonlinear terms in the incoming frequency  $\omega$  associated with the coefficients  $\alpha$  as well as  $\gamma_{\text{NLI}}$ , where NLI stands for nonlinear index, terms in  $2\omega$  associated with  $\beta_{\text{SHG}}$ , where SHG stands for second harmonic generation, and terms in  $3\omega$  associated with  $\gamma_{\text{THG}}$ , where THG stands for third harmonic generation, as well as many others when higher order nonlinear tensorial contributions are introduced.



**Fig. 17.3.** Two-level quantum system

This dipole can be interpreted as an oscillating electronic current on the molecular level, and the molecule itself as a receiving and emitting antenna whose anharmonicities, sometimes viewed as distortions in ‘normal’ linear propagation, play an important and positive role here. This differs from the mechanism of charge transport, in the sense used for example in semiconductors, by the fact that the electron charge carriers thereby driven oscillate more or less harmonically about an equilibrium position. The mechanical energy acquired from the exciting field is then restored in the form of radiation, emitted mainly at the same frequency (the case of the linear response induced by the polarisability  $\alpha$ , to be distinguished from the nonlinear effect due to the term  $\gamma_{\text{NLI}}$ ), but also at other frequencies (those of the various nonlinear mechanisms), which may in fact be favoured, as we shall see later in the context of classical and quantum considerations related to molecular engineering of optical nonlinearities.

The tensor  $\alpha$  expresses the harmonic response of the molecular system: irradiation at frequency  $\omega$  induces a Fourier component of the dipole at the same frequency  $\omega$ , which in turn radiates at this frequency. The real and imaginary parts of  $\alpha$  are related respectively to the phase shift and amplitude of the field radiated by this mechanism, i.e., to the refractive index and absorption of the molecular medium, respectively. Using the simple two-level quantum model shown in Fig. 17.3, direct application of first order time-dependent perturbation theory leads to the following expression for the linear polarisability:

$$\alpha \approx \frac{2\hbar\Omega_1\mu_{01}^2}{(\hbar\Omega_1)^2 - (\hbar\omega)^2 + \Gamma^2}, \quad (17.5)$$

where  $\Gamma$  represents the width of the spectral line of the excited state, or the reciprocal of the lifetime of the excited level. There are two contributions here, one arising from intrinsic properties of the molecule, which shows up through the square of the transition dipole  $\mu_{01} = \langle 0|\hat{\mu}|1\rangle$ , and a dispersive contribution with a resonant denominator with respect to the mismatch in the resonance between the photon energy and the energy of the electron transition. This first quantity couples the wave functions of the ground and excited states via the dipole observable (quantum equivalent of the classical dipole obtained by the principle of correspondence), which they enclose to give the corresponding vectorial matrix element.

It should be noted that, unlike the dipole of the ground state or the excited state, this quantity is not a physical observable constrained by strict symmetry and relatively ‘visual’ requirements, associated with a classically representable

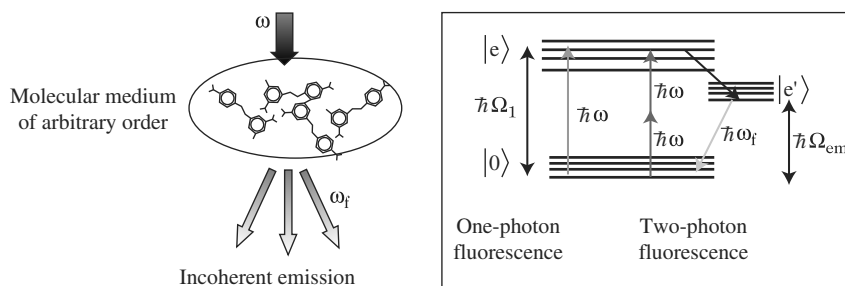


Fig. 17.4. One-photon and multiphoton fluorescence

mean value, such as the spatial moments of an electron charge density. A physical observable is defined here as the expected value of a physical quantity in a quantum state, e.g.,  $\mu_0 = \langle 0|\hat{\mu}|0\rangle$  or  $\mu_1 = \langle 1|\hat{\mu}|1\rangle$ ,  $\Delta\mu = \mu_1 - \mu_0$  for the dipole, or  $Q_0 = \langle 0|\hat{\mu} \otimes \hat{\mu}|0\rangle$  for the quadrupole, etc. More subtle symmetry requirements also affect non-observable quantities such as the above transition dipole and a systematic investigation must make use of group theory. However, simpler parity considerations may often lead in a straightforward manner to useful results in the form of selection rules, which express the existence or impossibility of optical transitions. They are based on the need for an even integrand in a spatial integral of type  $\langle 0|\hat{A}|1\rangle$ , which in turn imposes joint parity constraints on the operator  $\hat{A}$  and the wave functions enclosing it.

Hence, if the vector operator  $\hat{A}$  is odd, as happens with the dipole, the states  $|0\rangle$  and  $|1\rangle$  must have different parities, so that the integrand is even and the transition allowed. When  $\hbar\omega$  is close to  $\hbar\Omega_1$ , the so-called resonant situation,  $\alpha$  then tends to become purely imaginary and one is dealing with an absorption mechanism. This absorption increases as the difference  $\hbar\Omega_1 - \hbar\omega$  decreases and the modulus of the transition dipole  $\mu_{01}$  increases.

Another approach, in fact perfectly equivalent, considers the population of states  $|0\rangle$  and  $|1\rangle$  under the effect of the field  $\xi^\omega$ . This leads to the Fermi golden rule, which also brings out the central role played by the transition dipole in the transition probabilities of electrons under electromagnetic excitation. It is possible to relate this to the resonant polarisability introduced above. Within the same framework, one may also consider emission processes starting from an electron population previously excited to an energy state from which the electrons may fall back down, either directly or eventually via an indirect channel, to the ground state by emitting light. (We shall consider only spontaneous emission phenomena here.) This situation is illustrated in Fig. 17.4, which shows a sequence beginning with excitation by one-photon resonant absorption and followed by light emission from a configuration at lower energy than the one responsible for the absorption. The two excited states, i.e., the absorbing and emitting states, are connected by a very fast non-radiative transition which generally reflects a change of conformation, whence the distinction between absorption dipole and emission dipole.

The intensity of the light emitted by one-photon fluorescence is then given by

$$I_{\text{F}}^{1\text{ph}} \propto P_{\text{abs}}^{1\text{ph}} P_{\text{em}}^{1\text{ph}}, \quad (17.6)$$

where  $P_{\text{abs}}^{1\text{ph}} \propto |\boldsymbol{\mu}_{\text{abs}} \cdot \boldsymbol{\xi}^\omega|^2$  and  $P_{\text{em}}^{1\text{ph}} \propto |\boldsymbol{\mu}_{\text{em}} \cdot \mathbf{e}_{\text{F}}|^2$  represent the probabilities of one-photon absorption and emission. The vectors  $\boldsymbol{\xi}^\omega$  and  $\mathbf{e}_{\text{F}}$  denote the electric field and the unit vector in the direction of analysis of the fluorescence. The notation  $\propto$  indicates that we have omitted various physical constants and experimental factors of proportionality accounting for collecting efficiency, etc. These factors will be omitted throughout for greater clarity.

Consequently, for a statistical distribution of molecules, the fluorescence intensity can be rewritten in the form

$$I_{\text{F}}^{1\text{ph}} \propto (\mathcal{I}^\omega \otimes \tilde{\mathbf{e}}_{\text{F}}) \cdot \int \alpha_{\text{abs}}(\Omega) \otimes \alpha_{\text{em}}(\Omega) F(\Omega) d\Omega, \quad (17.7)$$

where  $\mathcal{I}^\omega$  and  $\tilde{\mathbf{e}}_{\text{F}}$  are rank 2 tensors (analogs of matrices) given respectively in terms of the electric field vectors  $\boldsymbol{\xi}^\omega$  and  $\mathbf{e}_{\text{F}}$ , by  $\mathcal{I}^\omega = \boldsymbol{\xi}^\omega \otimes \boldsymbol{\xi}^\omega$  and  $\tilde{\mathbf{e}}_{\text{F}} = \mathbf{e}_{\text{F}} \otimes \mathbf{e}_{\text{F}}$ . The tensor  $\alpha_{\text{abs}}$  is given by the two-level expression (17.5), where the dipole matrix element is taken between the levels relevant to the resonant absorption, i.e.,  $|0\rangle$  and  $|e\rangle$ . Likewise, the transition dipole relating to  $\alpha_{\text{em}}$  is considered between the relevant fluorescence levels, i.e.,  $|e'\rangle$  and  $|0\rangle$ . A key point concerning symmetries is the total contraction operation, i.e., summation over all four common Cartesian indices, denoted by  $\cdot$  in (17.7), which expresses the coupling between the tensor product of the incident and readout field tensors and the corresponding product of the polarisability tensors relating to absorption and emission (in that order).

These considerations generalise to another configuration, depicted in Fig. 17.9, which is of great importance in biophotonics and particularly in confocal microscopy. This describes the fluorescence excitation, still in terms of the same states  $|e'\rangle$  for emission and  $|e\rangle$  for absorption, by a two-photon absorption mechanism. This mechanism was predicted almost at the beginning of quantum mechanics by M. Göppert-Mayer, who subsequently became famous for the liquid-drop model of nuclear structure. However, it was only actually observed some thirty years later, since this kind of experiment requires pumping powers that only became available with lasers. Quite naturally, the name of M. Göppert-Mayer was later used to name the unit of the two-photon absorption cross-section.

In an analogous way to one-photon fluorescence, the intensity of two-photon fluorescence can be written  $I_{\text{F}}^{2\text{ph}} \propto P_{\text{abs}}^{2\text{ph}} P_{\text{em}}^{1\text{ph}}$ , where  $P_{\text{abs}}^{2\text{ph}} = |\boldsymbol{\mu}_{\text{abs}} \cdot \boldsymbol{\xi}^\omega|^4$  for the probability of two-photon absorption, and  $P_{\text{em}}^{1\text{ph}} = |\boldsymbol{\mu}_{\text{em}} \cdot \mathbf{e}_{\text{F}}|^2$  for the probability of one-photon emission from the fluorescence state. Hence, for a statistical distribution of molecules, we find

$$I_{\text{F}}^{2\text{ph}} \propto (\mathcal{I}^\omega \otimes \mathcal{I}^\omega \otimes \tilde{\mathbf{e}}_{\text{F}}) \cdot \int_{\Omega} \gamma_{\text{abs}}^{2\text{ph}}(\Omega) \otimes \alpha_{\text{em}}(\Omega) F(\Omega) d\Omega. \quad (17.8)$$

The tensorial term  $\gamma_{\text{abs}}^{2\text{ph}}$  (purely imaginary for resonant absorption, which is almost always the case) combined with the tensor  $\mathcal{I}^\omega \otimes \mathcal{I}^\omega$  quadratic in the field expresses the two-photon absorption and arises in the expression for the induced dipole given earlier.

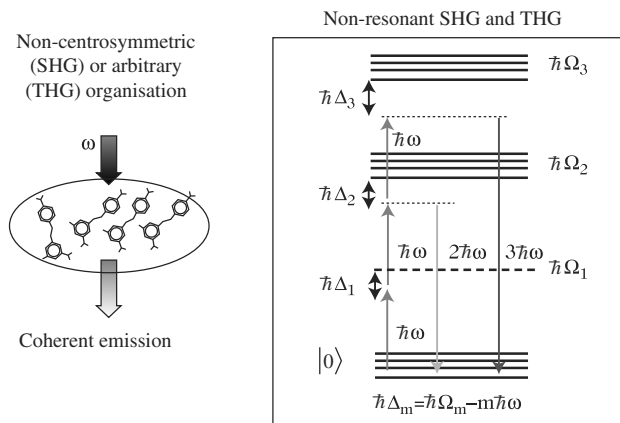
Its quantum expression in the two-level model can be obtained from a higher order perturbative calculation (an order 3 perturbation in the field) than the one leading to  $\alpha$ . This leads to the expression given at the end of this section.

The emission of fluorescence by absorption of one or two photons plays an important role in biophotonics on the microscale, and more recently, on the nanoscale, as we shall see in more detail in Sects. 17.2 and 17.3. The introduction of fluorescent markers able to attach themselves selectively to some particular cellular compartment or physicochemical medium has made it possible to detect the presence and spatial distribution of targeted media, and also to follow them in time by microscopy techniques which today allow resolutions of the order of a single molecule at in-vivo temperatures.

The fluorescence effects discussed above are not the only manifestations of photonics able to inform us about some biological context. More recently, the Raman effect related to variations in electron polarisability resulting from coupling between electron excitation and vibrational modes has been used for nanoscale monitoring of molecular species that can be identified through spectral shifts induced by these modes and referred to by their Stokes or anti-Stokes sign.

The use of nonlinear effects in the strict sense is a more recent achievement, based on the direct reemission by the induced dipole in the form of  $2\omega$  or  $3\omega$  radiation corresponding to the appropriate terms in  $\beta$  and  $\gamma$  in the dipole expansion (17.2). In the quantum level diagram of Fig. 17.5, the downward arrows of amplitude  $2\hbar\omega$  and  $3\hbar\omega$  indicate the instantaneous restitution, without spectral shift, i.e., without passing through some third state via a non-radiative, phase-shifting transition, of the energy absorbed by virtual absorption of two or three photons, respectively.

The notion of virtual absorption, as opposed to the classic absorption effects shown in Fig. 17.5 where the level at  $2\hbar\omega$  is situated in the absorption band of the medium, relates to the magnitude of the energy shift  $\Delta_m = \hbar\Omega_m - m\hbar\omega$  ( $m = 1, 2, 3$ ) between one of the excited levels at energy  $\hbar\Omega_m$  and the total energy  $m\hbar\omega$  of the  $m$  photons at  $\omega$  that cascadingly interact with the medium through a multiphoton absorption process. In the case of a two-level system,  $\hbar\Omega_m$  can be defined as the bottom of the absorption band or the position (referred to as the  $\lambda_{\text{max}}$  position) corresponding to the absorption maximum. In either case, we speak of a virtual transition when  $\Delta_m$  is such that the  $m$ -photon transition does not induce a significant population of the level  $\hbar\Omega_m$ . This can be quantified by the Fermi golden rule, extended by perturbations to the multiphoton process, which serves to evaluate the corresponding transition probability. Such a probability is rather small in comparison with an observational threshold that must be fixed in consequence of the experimental setup and detector sensitivity.



**Fig. 17.5.** Generation of harmonics

In such a configuration, the molecules behave as anharmonic nano-oscillators, absorbing and emitting light at frequencies  $2\omega$  and  $3\omega$ , due respectively to the cubic and quartic anharmonicities of the potential felt by the electrons as they are being driven by the incoming laser field at  $\omega$ . There is no spectral shift as in the fluorescence mechanisms already described, but there may be dissipation, i.e., a reduction in the conversion efficiency of the incident intensity at  $\omega$  into the emitted intensity at  $m\omega$ . This occurs when the parameter  $\Delta_m$  expressing the deviation from resonance becomes small. One then moves from a virtual absorption scenario, i.e., without dissipation, in the absence of a residency lifetime related to an effective population in the excited state, to a real absorption scenario for which the lifetime of the excited state, related to its spectral line width, reflects such dissipation (in the classic form of a first order friction term in the dynamical equation of the polarisable electron). The intensities emitted at  $2\omega$  and  $3\omega$  can be expressed in almost the same formalism as the one already used in the fluorescence discussion, but with a significant difference due to the appearance of a coherent contribution. In contrast to the fluorescence phenomena, for which non-radiative relaxation mechanisms entail at least a partial loss of phase memory and incoherent emission, the generation of harmonics considered here rests upon virtual transitions with negligible phase shift. The emission from a set of  $N$  dipoles, assumed to be contained in unit volume, then leads to a radiated intensity given by

$$I^{m\omega} \propto \left( \sum_i^S \mu_i^{m\omega} \right) \cdot \left( \sum_j^S \mu_j^{m\omega} \right)^* , \quad (17.9)$$

where sums are taken over individual molecules indexed by  $i$  or  $j$ , and the asterisk indicates the operation of complex conjugation. This intensity can

then be decomposed into two fundamentally distinct contributions:

$$I^{m\omega} = I_{\text{incoh}}^{m\omega} + I_{\text{coh}}^{m\omega} \propto \sum_i^S |\boldsymbol{\mu}_i^{m\omega}|^2 + \left| \sum_j^S \boldsymbol{\mu}_j^{m\omega} \right|^2. \quad (17.10)$$

The first sum corresponds to incoherent effects or to terms with the same index  $i = j$  in the double sum of (17.9), whose effects add in amplitude. Indeed, whatever the phase correlation of the dipoles, it does not play a role in the term  $I_{\text{incoh}}^{m\omega}$ , given that the corresponding term  $|\boldsymbol{\mu}_{i=j}^{m\omega}|^2$  is real. The second term corresponds to a coherent sum of cross terms  $\boldsymbol{\mu}_i^{m\omega} \cdot \boldsymbol{\mu}_j^{m\omega*}$  for  $i \neq j$ . This term is nonzero insofar as the modulus of the vector sum  $\sum_j^S \boldsymbol{\mu}_j^{m\omega}$  is not itself zero. This imposes strong symmetry constraints on the array of dipoles and in particular forbids a centrosymmetric arrangement, in contrast to the incoherent contribution, made up as it is of a sum of squared moduli of vectors which can never be zero. The sums  $\sum^S$  can also be considered as statistical sums over random angle variables insofar as the ergodicity principle is valid. Taking the limit and the average over the angle variables, one arrives in a similar manner at the radiated intensity:

$$I^{m\omega} \propto N \int_{\Omega} |\boldsymbol{\mu}^{m\omega}(\Omega)|^2 F(\Omega) d\Omega + N^2 \left| \int_{\Omega} \boldsymbol{\mu}^{m\omega}(\Omega) F(\Omega) d\Omega \right|^2. \quad (17.11)$$

Then introducing the perturbation expansion (17.2) of the dipole, the coherent and incoherent contributions can be written in intrinsic tensorial form as

$$\begin{aligned} I_{\text{coh}}^{2\omega} &\propto N^2 \tilde{e}^{2\omega} \otimes \mathcal{I}^\omega \otimes \mathcal{I}^{\omega*} \cdot \int \beta(\Omega) F(\Omega) d\Omega \otimes \int \beta^*(\Omega) F(\Omega) d\Omega, \\ I_{\text{coh}}^{3\omega} &\propto N^2 \tilde{e}^{3\omega} \otimes \mathcal{J}^\omega \otimes \mathcal{J}^{\omega*} \cdot \int \gamma(\Omega) F(\Omega) d\Omega \otimes \int \gamma^*(\Omega) F(\Omega) d\Omega, \\ I_{\text{incoh}}^{2\omega} &\propto N \tilde{e}^{2\omega} \otimes \mathcal{I}^\omega \otimes \mathcal{I}^{\omega*} \cdot \int \beta(\Omega) \otimes \beta^*(\Omega) F(\Omega) d\Omega, \\ I_{\text{incoh}}^{3\omega} &\propto N \tilde{e}^{3\omega} \otimes \mathcal{J}^\omega \otimes \mathcal{J}^{\omega*} \cdot \int \gamma(\Omega) \otimes \gamma^*(\Omega) F(\Omega) d\Omega, \end{aligned} \quad (17.12)$$

where  $\tilde{e}^{m\omega} = \mathbf{e}^{m\omega} \otimes \mathbf{e}^{m\omega}$ ,  $\mathcal{I}^\omega$  (rank 2 tensor) as before, and  $\mathcal{J}^\omega = \boldsymbol{\xi}^\omega \otimes \boldsymbol{\xi}^\omega \otimes \boldsymbol{\xi}^\omega$  is a new tensor of rank 3 relating to three-photon processes.

The statistical sum over the angular variable  $\Omega$  becomes a discrete sum over the unit cell of the lattice in the case of a crystal (where molecules have a fixed orientation in the crystal lattice). For molecules oriented in an electric field in a polymer,  $F(\Omega)$  is generally given by a Maxwell–Boltzmann distribution reflecting the orientation mode of the medium. A conceptually and practically important example is the Langevin-type process where molecules are oriented in a field by coupling between the permanent dipole  $\boldsymbol{\mu}_0$  of the elementary molecular entities and a static external electric field  $\mathbf{E}_0$ . In the typical case of one-dimensional rod-shaped molecules, the medium in a field exhibits a cylindrical statistical symmetry and the three Euler angles then reduce to a single azimuthal angle  $\theta$ , measured from the axis of the orienting field. The statistical distribution (probability distribution of the molecular orientations in a unit volume) is given in this case by

$$F(\Omega) = f(\theta) \propto \exp\left(-\frac{\mu_0 E_0 \cos \theta}{kT}\right).$$

This can then be approximated by the first order term in  $E_0$ , which simplifies calculation of the average values of molecular observables and leads to analytical expressions that can be interpreted physically. This field of investigation is attracting great interest at the moment, largely to satisfy the strategic requirements of electro-optical components for optical telecommunications purposes, whilst the underlying physics is beginning to find applications in biophotonics, e.g., in cases where  $E_0$  corresponds to an electrophysiological potential, such as in neurones or cellular membranes.

The tensors  $\beta$  and  $\gamma$  occurring in the above expressions for the intensities of the second and third harmonics are given in the context of a two-level quantum model by

$$\beta_{\text{SHG}} = \frac{3\hbar\Omega_2^2\mu_{02}^2\Delta\mu^2}{2\left[\hbar^2\Omega_2^2 - (2\hbar\omega)^2 + \Gamma^2\right](\hbar^2\Omega_2^2 - \hbar^2\omega^2)} \quad (17.13)$$

and

$$\gamma_{\text{THG}} = \mu_{03}^4 G^{\text{THG}}(\omega, \Omega_3) - \mu_{03}^2 \Delta\mu_3^2 F^{\text{THG}}(\omega, \Omega_3), \quad (17.14)$$

where the various contributions are defined above and in Fig. 17.5.  $\Delta\mu_2 = \mu_2 - \mu_0$  is the difference between the permanent dipoles of the relevant ground states, and likewise for  $\Delta\mu_3 = \mu_3 - \mu_0$ . Likewise  $\mu_{02}$  (resp.  $\mu_{03}$ ) is the transition dipole between the ground state and state  $|2\rangle$  (resp.  $|3\rangle$ ). The dispersion factors  $G^{\text{THG}}(\omega, \Omega_3)$  and  $F^{\text{THG}}(\omega, \Omega_3)$  are given by the resonant expressions when  $\Delta_3$  tends to zero. A precise formulation can be found in the specialised literature.

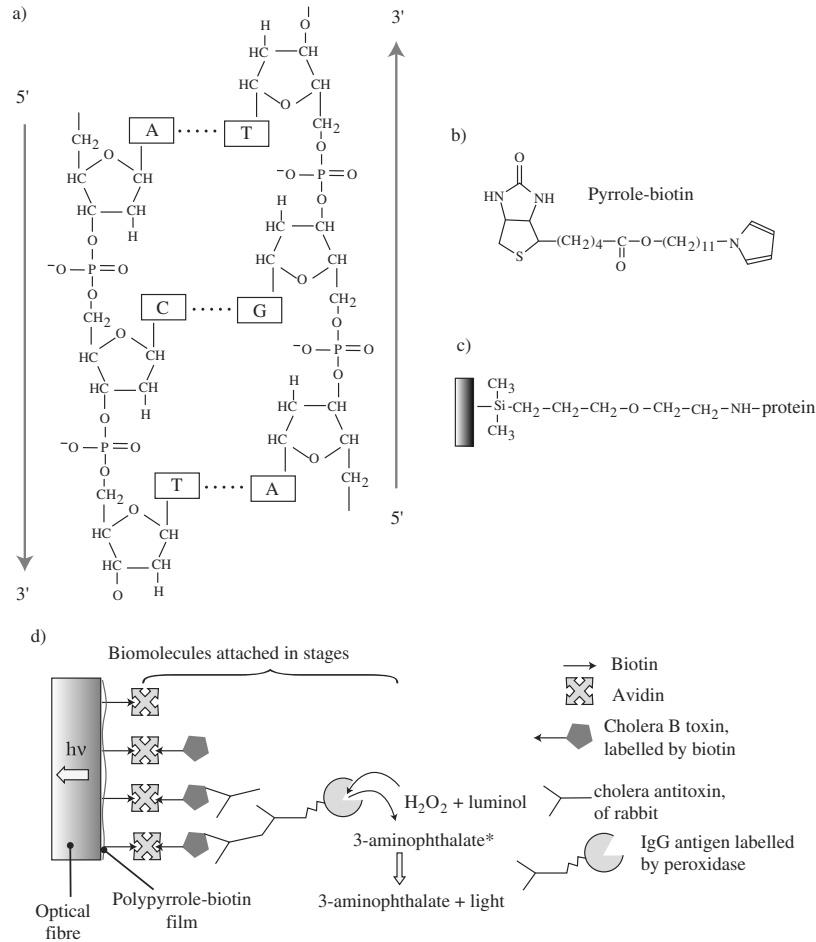
## 17.2 Molecules, Supramolecular Assemblies, and Nanoparticles

### 17.2.1 Coupling Between Nanoparticles and Biomolecules

The association of biomolecules and nano-objects with photonic properties like those defined in the last section is a major issue in nanobiophotonics and most of its future developments are expected to be reached by this route. These new architectures are based on the characteristic associative properties of biologically relevant macromolecules. In this context, we may distinguish three main types of associative principle likely to contribute to the hybridisation of nanometric organic and inorganic entities with biological entities, but also to the association of such entities with one another (see Fig. 17.6):

- Interactions of type receptor–ligand, one of the most common being the streptavidin–biotin pair, based on a particularly efficient and selective lock-and-key affinity.





**Fig. 17.6.** Hybridisation of biological and nanometric inorganic entities. (a) Specific matching of strands by hydrogen bridging between base pairs. Selectivity of G-C and A-T coupling allows specific recognition of sequences of a DNA strand by the so-called complementary sequences of an adjacent fragment. It constitutes the underlying principle of the DNA chip used in post-genome research and diagnostic genetics, and also the basis of replication in life. (b) Biotin molecule modified by addition of a pyrrole group, which allows complexation via the sulfur atom, on a metallic (Au, Ag) surface. (c) Surface silanisation is used to attach biologically relevant molecules such as proteins to surfaces by strong covalent bonding. (d) The biotin molecule associates strongly and highly selectively with one of the four binding sites of a protein known as avidin (or its variant streptavidin). The example chosen is a fibre optic sensor which detects the autoluminescence collected by the fibre from a cholera antitoxin. The toxin-antitoxin matching illustrates the specificity of antigen-antibody recognition [21]. Moreover, an ITO layer (transparent electrode) has in this case been intercalated between the silica and the polypyrrole layer (thickness  $200\ \mu\text{m}$ )

- Antigen–antibody interactions.
- The complementarity between pairs of bases making up fragments of DNA strands (also called nucleotides).

Such associations form a genuine toolbox in nanobiophotonics, making it possible to associate a wide range of nanocrystalline structures with practically any biologically important molecule or macromolecule. The nanocrystalline structures in question often display photophysical properties far superior to those of organic dyes, which are sometimes inadequately resistant to the effects of prolonged or intense light radiation.

Apart from the problem of tethering functional entities, in particular, light-emitting, but also magnetic, or even photothermic entities, a key issue is the self-organisation of nano-objects with typical sizes between a few nanometers and a few hundred nanometers. Indeed, this size range cannot be reached by top-down approaches, which come up against the physical limits imposed on microlithography by the laws of diffraction at visible and near-UV frequencies. On the other hand, the complementary bottom-up approach, which proceeds by conventional supramolecular chemistry, is able to reach the 5–10 nm range and below.

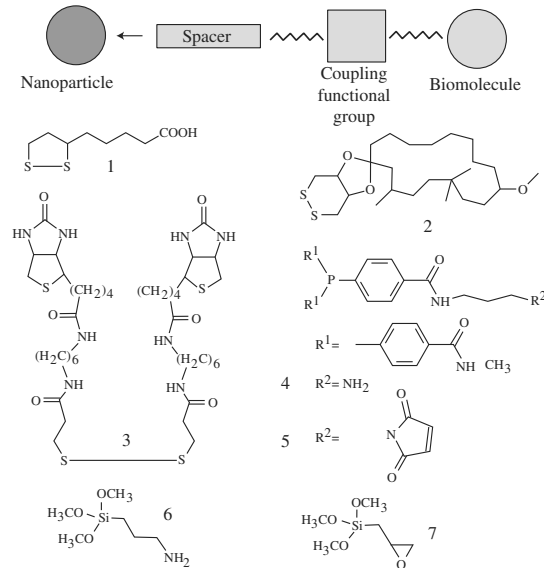
Figure 17.7 shows the basic structure of an entity which conjugates a biomolecule with a metallic (e.g., gold or silver) nanoparticle with the help of a cross-linking agent and a functional group, several examples of which are given in Table 17.3.

As a rule the interactions associating nanoparticles with biomolecules are electrostatic and non-covalent. The thiol group thus exhibits a strong affinity for gold surfaces and the presence of two thiol groups rather than one, as is the case for compounds 1 and 3 in Fig. 17.7 strengthens this cohesion. The adsorption of proteins stabilises the surface of the nanoparticle and opposes flocculation which would otherwise tend to aggregate and precipitate the nanoparticles. The strong complementarity of the streptavidin–biotin pair also plays an essential role in the associative link.

### Nanoparticle Organisation Mediated by Biomolecules

The equivalent on the nanoscale of genuine supramolecular constructions, as well as aggregates with controlled structures in which the basic building blocks are not atoms but nanostructures can be fabricated by associating nanoparticles with the help of biological molecules whose mutual recognition properties are thereby put to use. This is illustrated in Fig. 17.8, which shows simple ‘nanomolecules’ made from identical or different nanoparticles. In the first case, at the two ends of the chain are two dinitrophenyl (DNP) groups in meta configuration. These are recognised and efficiently complexed by immunoglobulin-type receptors. This leads to a homodimer associated by the symmetric organic group.

By asymmetrising the linker molecule provided with a D-biotin-type molecule at one end and the same dinitrophenyl group as before at the other end,



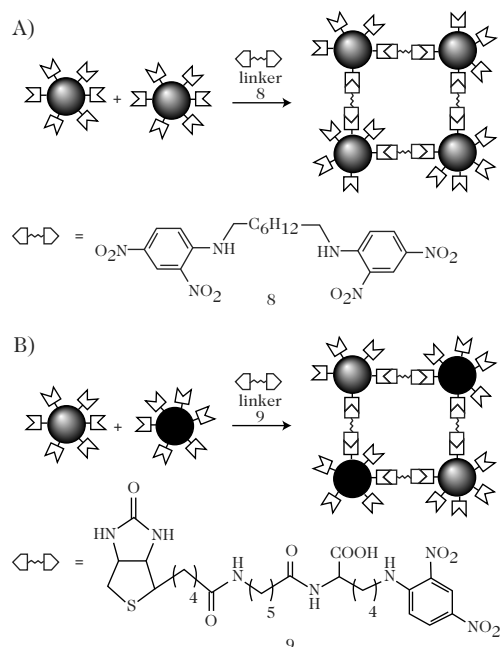
**Fig. 17.7.** Coupling of a biomolecule with a metallic nanoparticle. From [22]

**Table 17.3.** Examples of coupling between a biomolecule and a metallic nanoparticle by virtue of a functional group and a cross-linking agent

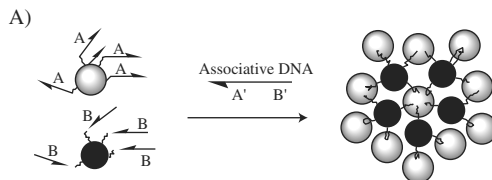
Nanoparticle	Cross-linking agent	Functional group	Biological molecule
Gold	Streptavidin	Biotin-(CH <sub>2</sub> ) <sub>6</sub>	DNA, immunoglobulins, albumin serum
Silver	Citrate	H <sub>2</sub> N-lysine	Heme protein, immunoglobulins
CdSe/ZnS		HS-(CH <sub>2</sub> ) <sub>6</sub>	DNA
CdSe/ZnS	HS-CH <sub>2</sub> -COOH and 1	H <sub>2</sub> N-lysine	Immunoglobulin, proteins
GaAs/InP	Aminated phosphoramidate	HOOC-glucose	Proteins

two different nanoparticles can be matched by previously coating one of them with an immunoglobulin that is a biotin receptor and the other with an immunoglobulin able to recognise DNP (called IgG). The matching continues beyond the dimer and it can be directly established by transmission electron microscopy that initially distant entities will aggregate as the suspension is enriched with the relevant cross-linking agent.

Note also that cross-linking agents of oligopeptide type specifically associating proteins with semiconductor surfaces such as GaAs (100 or 111), InP (100) and Si(100) can be designed and implemented.



**Fig. 17.8.** Nanoparticles assembled using biological macromolecules. From [22]



**Fig. 17.9.** Assembly using DNA. From [22]

A particularly spectacular way of assembling nanoparticles uses the complementarity of DNA strands grafted onto different nanoparticles. The latter thus develop a strong tendency to form specific mutual associations. One advantage of these constructions is the very good calibration of distances between nanoparticles, due to the rigidity of the DNA strands which do behave here very much like calibrated spacers. Furthermore, nature provides a complete range of biomolecular reagents such as the endonucleases, ligases and various other enzymes that can be used to process, e.g., segment or hybridise, DNA-based constructions with the highest level of specificity and precision of the order of 1 Å.

A general approach to the realisation of such constructions is shown in Fig. 17.9. Depicted are two non-complementary oligonucleotides which functionalise the surface of two different nanoparticles. The corresponding

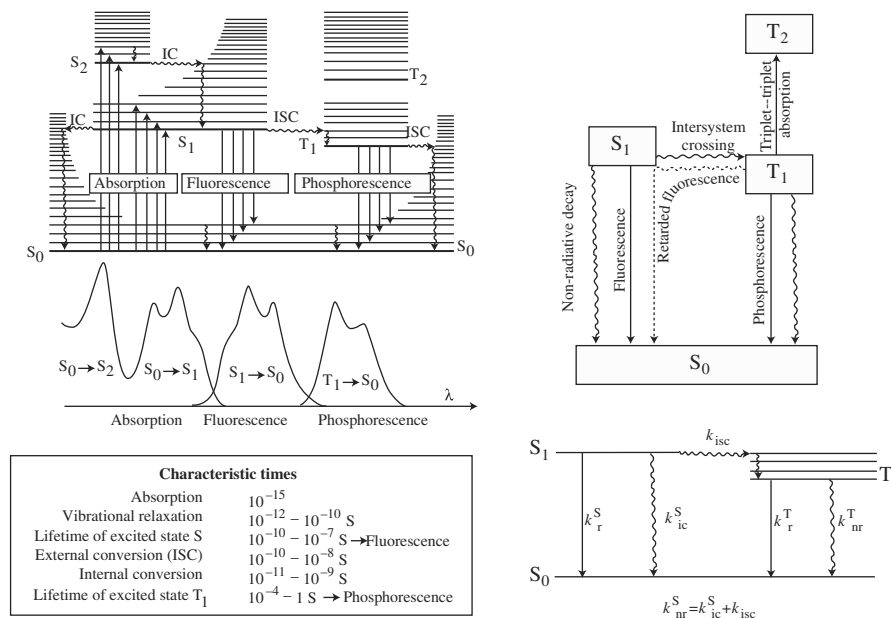
cross-linking agent comprises three parts: a central part in the form of a highly cohesive double helix terminated by two individual strands complementary to the oligonucleotides covering the nanoparticle surfaces. When this cross-linking agent is added, the nanoparticles match up and arrange themselves into a well-ordered 3D lattice. In this way, nanostructures of type CdSe/ZnS (discussed further below) have been associated with gold nanoparticles, whence they have been arranged into well-structured aggregates of dimers. Cooperative electronic effects have been demonstrated in these structures using fluorescence. This result exemplifies a broader future avenue for material engineering, in which biomolecules act as templating agents, with a view to creating new generations of biomimetic materials.

In the same way, one may also combine the specific matching properties of the base pairs in DNA fragments with the strong affinity of the biotin–streptavidin (STV) pair. The STV–DNA pair provides four native complexation sites for the biotin molecules, themselves attached to a gold nanostructure. Hence, the possibility of forming a tetrahedral cluster of such nanostructures. One may then play on the affinity by complementarity of the bases on the DNA strand conjugate to STV to attach these constructions on a substrate that has been previously functionalised by complementary oligonucleotides, or to attach an immunoglobulin that can then be used to target in a specific manner tissues or other types of biological substrate, for the purposes of diagnostic studies, among other possibilities relating to molecular electronics.

### Fluorescence as Biological Marker

The fluorescence of artificial or natural molecules is a well known phenomenon that biologists have long sought to exploit as a way of marking biological material (tissue, cells, proteins, macromolecular complexes like viruses, etc.), both *in vivo* and *in vitro*. The idea is then to monitor the evolution of these tagged systems by one or several suitable techniques, e.g., one-photon or multiphoton confocal microscopy, possibly associated with second or third harmonic generation. The aim is to induce from these observations, e.g., of the spectral drift and/or polarization features of the emission, relevant information concerning the environment or internal modifications of the system.

The ideal system must therefore fulfill a certain number of requirements. One is efficiency or yield, i.e., the ability to generate a maximal optical signal for a minimum of incident energy, and this not only to respond to the always legitimate concern of optimising the energy conversion yield, but also quite simply to avoid damaging tissues that are by their very nature fragile, or again to avoid exciting interference phenomena that can mask the targeted fluorescence signal. Markers must also be very small so that they do not interfere with the real system by artificial and adverse steric effects, and they must have as long a lifetime as possible (in particular under optical excitation). Finally, they must be chemically stable and be able to associate in a controllable manner with a predetermined site of the relevant system.



**Fig. 17.10.** Absorption and emission processes. See text for explanations and also [23]

No system will readily satisfy all these criteria in a completely satisfactory way. They are demanding in themselves and even more difficult to reconcile as an ensemble. In this context, fluorescent molecules are justifiably reckoned to be good candidates, although their lifetime often remains something of a problem. They give rise to bleaching phenomena under exposure to visible light for times that rarely exceed a few hours or even a few minutes, but are most commonly of the order of several tens of seconds (see Sect. 17.3 for a discussion of individual molecules). Figure 17.10 and Table 17.4 review the basic mechanisms involved in emission and absorption of light and the associated orders of magnitude for the prototypical family of polycondensed aromatic hydrocarbons. The different transition rates involved in the light-molecule exchanges are indicated in Fig. 17.10. The orders of magnitude of the lifetimes derived from the reciprocals of these rates are also indicated.

The constant  $k_r^S$  is the rate (number of events per second) of radiative decay from the state  $S_1$  to the state  $S_0$  accompanying the fluorescence emission. The internal conversion rate  $k_{ic}^S$  expresses the rate of return to equilibrium of the population of the excited state by non-radiative transition under the effects of structural reorganisation or energy exchange, e.g., by collision with surrounding solvent molecules if the experiment occurs in solution.

The constant  $k_{isc}$  is the intersystem conversion rate which expresses the transition from the excited singlet state  $S_1$  to the triplet state  $T_1$ , possibly

**Table 17.4.** Fluorescence in hydrocarbons. Notation is explained in the text. From [23]

Compound	Solvent (temperature)	$\phi_F$	$\tau_S$ [ns]	$\phi_{lec}$	$\phi_P$	$\tau_T$ [s]
Benzene	Ethanol (293 K)	0.4	31			
	EPA (77 K)				0.17	7.0
Naphthalene	Ethanol (293 K)	0.21	2.7	0.79		
	Cyclohexane (293 K)	0.19	96			
	EPA (77 K)				0.06	2.6
Anthracene	Ethanol (293 K)	0.27	5.1	0.72		
	Cyclohexane (293 K)	0.30	2.24			0.09
	EPA (77 K)					
Perylene	n-Hexane	0.98		0.02		
	Cyclohexane (293 K)	0.98	6			
Pyrene	Ethanol (293 K)	0.65	410	0.35		
	Cyclohexane (293 K)	0.65	450			
Phenanthrene	Ethanol (293 K)	0.13		0.85		
	n-Heptane (293 K)	0.16	0.60			
	EPA (77 K)				0.31	3.3
	Polymer film	0.12		0.88		0.11

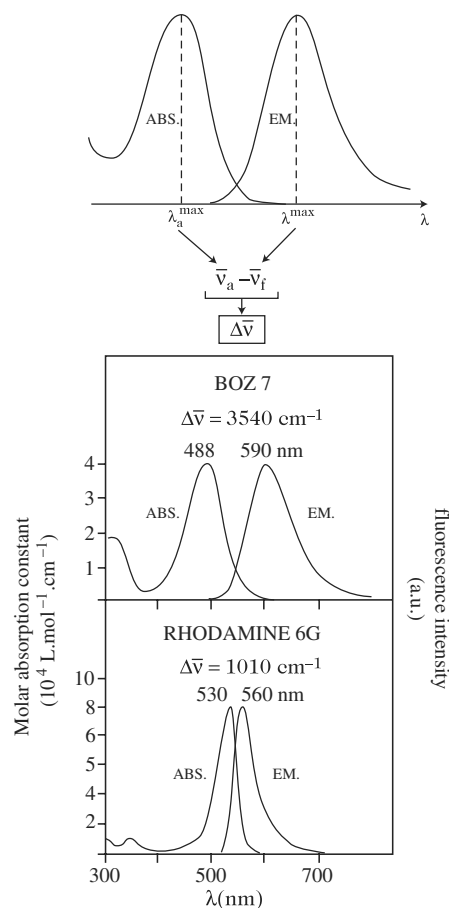
followed by emission of so-called phosphorescence radiation, which is only weakly allowed and hence retarded with respect to fluorescence. Although the latter manifests itself at times less than the nanosecond, and even down to the picosecond, phosphorescence is generally bound to occur on a much longer characteristic time scale ranging from the microsecond to the second. The radiative and non-radiative deactivation constants of the triplet state towards the singlet ground state are denoted  $k_r^T$  and  $k_{nr}^T$ , respectively.

By analogy with chemical kinetics, the decay rate of excited molecules is given by a population rate equation of the form

$$-\frac{dN_1}{dt} = (k_r^S + k_{nr}^S) N_1, \quad (17.15)$$

where  $k_{nr}^S = k_{ic}^S + k_{isc}$  and  $N_1$  is the number of molecules in the excited state. With this notation the characteristic decay time for the excited state  $S_1$  is given by  $\tau_S = (k_r^S + k_{nr}^S)^{-1}$  whilst that of the triplet state is given by  $\tau_T = (k_r^T + k_{nr}^T)^{-1}$ . The fluorescence quantum yield, which may be close to unity for the most efficient molecules, e.g., perylene in Table 17.4, is the fraction of the population of excited molecules to return to the ground state by a radiative channel with emission of a fluorescence photon, viz.,

$$\Phi_F = \frac{k_r^S}{k_r^S + k_{nr}^S} = k_r^S \tau_S. \quad (17.16)$$



**Fig. 17.11.** Stokes shift for fluorescence emission. From [23]

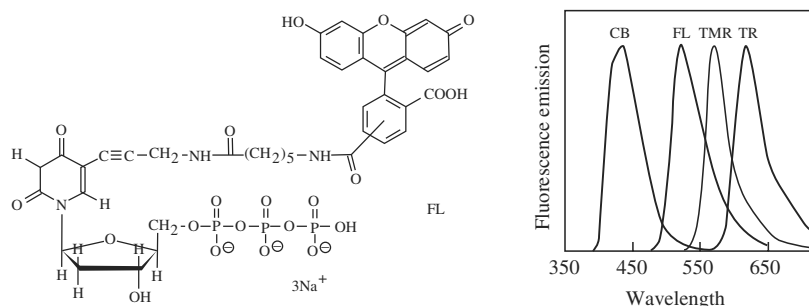
Likewise, the quantum efficiencies of intersystem conversion and phosphorescence,  $\Phi_{\text{isc}}$  and  $\Phi_{\text{p}}$ , respectively, are given by

$$\Phi_{\text{isc}} = k_{\text{isc}}\tau_{\text{S}} , \quad \Phi_{\text{p}} = \frac{k_{\text{r}}^{\text{T}}}{k_{\text{l}}^{\text{T}} + k_{\text{nr}}^{\text{T}}} \Phi_{\text{isc}} . \quad (17.17)$$

A particularly important parameter from a practical standpoint, in order to distinguish the fluorescence emission from the exciting beam, is the Stokes shift. This expresses the shift between the maximum of the first absorption band and the maximum of the fluorescence spectrum, which one thus seeks to maximise, the amplitude of this shift increasing with the polarity of the solvent. Fluorescence thus provides information about the latter. Figure 17.11 shows the Stokes shift for two fluorescent molecules: a derivative of benzoxanone and the standard dye Rhodamine 6G.

The technique known as FISH (fluorescence in situ hybridisation) involves attaching fluorescent molecules onto oligonucleotides. This approach





**Fig. 17.12.** A dye (FL) grafted onto a DNA strand (*left*) and emission spectra of various dyes (CB, FL, TMR, TR) (*right*). From [15]

is tending to replace radioactive labelling as a way of tracing DNA or proteins and monitoring their motion in situ. A strand of test DNA is made fluorescent by adjoining a fluorescent part that is discrete enough and suitably well located not to interfere with hybridisation. The modified DNA strand is designed to match up with the specific complementary sequence of a human or other chromosome. Figure 17.12 shows a fluorescein molecule grafted onto a base (thymine), itself connected to the characteristic sucrose–phosphate sequence of a DNA strand. Other dyes with quite distinct emission spectra can be used for specific grafts. One can then achieve multiplexed wavelength marking, e.g., where each colour characterises one of the 23 human chromosomes, making them individually identifiable.

### 17.2.2 Luminescent Nanostructures Based on Semiconductors and Metals

When it comes to marking biochemical species, some of the disadvantages displayed by fluorescent molecules can be removed by replacing them by inorganic nanostructures, the most widely used being based on II–VI direct gap semiconductors such as ZnS, CdSe, or some of their ternary or quaternary alloys. The main disadvantages at issue here are a certain fragility, as well as broad emission spectra which generally make it impossible to employ more than a relatively small number of markers (at most 3 or 4), thus reducing the combinatorial potential of this method.

The quasi-atomic nature of the optical transitions involved in the mechanisms for absorption and emission of light in confined structures are indeed capable of leading to much narrower line widths. What is more, these widths are strongly dependent on the size of the nano-object and can be effectively tuned. The typical dimension of such nano-objects is of the order of the natural radius of the electron–hole pair (exciton) created in the semiconductor, or even smaller. This results in useful phenomena known as quantum confinement effects, characteristic of geometries in which the wave function of the

elementary excitation ‘feels’ the boundary conditions imposed by the walls of the nanostructure. The typical dimensions of these phenomena are of the order of a few nanometers.

The science of materials has long been actively engaged in establishing efficient fabrication methods that can produce nanoparticles with well-calibrated sizes, largely for physical purposes. These procedures may advantageously use soft organometallic chemistry (in wet phase), the products of which are then refined by a size-selective precipitation, resulting in monodispersed nanostructures with a standard deviation as low as a few percent in the size distribution.

A very simple model (see Fig. 17.13), which was subsequently refined in various specific ways, still has a general validity. It can be used to relate the size variation in the nanostructure to the energies of the excited levels. Assuming a spherical shape, the latter are indexed by the same radial and orbital quantum numbers ( $n$  and  $l, m$ , respectively) as the atomic orbitals (whence the term ‘quasi-atom’ sometimes used to designate such nano-objects):

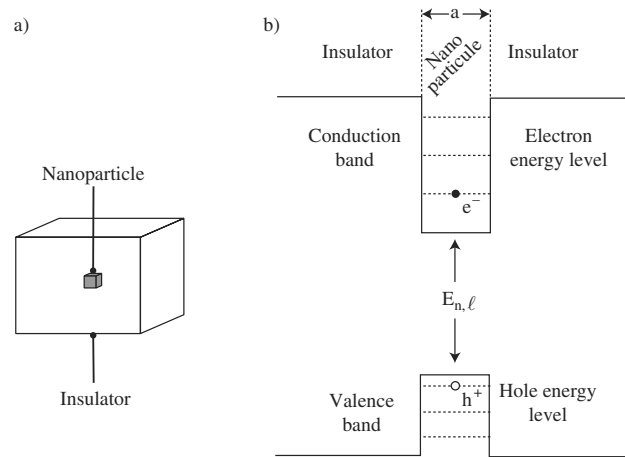
$$E_{n,l} = \frac{\hbar^2 \alpha_{n,l}^2}{2m_0 a^2}, \quad (17.18)$$

where  $\alpha_{n,l}$  is the  $n$ th zero of the spherical Bessel function of order  $l$ ,  $m_0$  is the electron mass, and  $a$  is the characteristic dimension of the nanostructure, assumed to be described here by an infinitely deep spherical potential well of the same radius. The appearance of the  $1/a^2$  term reflects a strong size dependence.

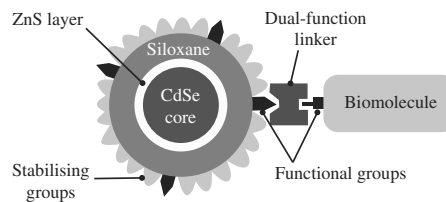
By controlling this crucial geometric parameter, one gains access to a whole range of light-emitting nanolabels made from the same semiconductor, viz., CdSe, emitting at distinct wavelengths for the same excitation wavelength depending on the size. Furthermore, they can be attached to a biological molecule or indeed a solid substrate. A prototype for such structures is shown schematically in Fig. 17.14. Surrounding the spherical core structure made from CdSe is an extra shell layer of ZnS, a material with larger gap than CdSe. This means that, from the standpoint of an electron, it represents a deeper potential well which further increases the level of confinement. The result is that the luminescence yield from such core-shell nanoparticles can be raised from 10% to more than 50%.

A further advantage obtained by using a single excitation wavelength is the possibility of marking many chosen places in a system by a whole set of nanoparticles. This allows the colocalised detection of several interacting entities, within a cell medium, for example.

Another problem is then the solubility, and more generally, the acceptability in the biological (i.e., aqueous) medium of an initially hydrophobic structure, not to mention its potential for grafting onto a biological entity. Several approaches have recently been proposed, e.g., coating by an outer layer of siloxane. This is a sol-gel hybrid polymer well-suited to soft chemistry in solution and capable of carrying organic functional groups like amines, thiols,



**Fig. 17.13.** Fluorescence emission in semiconductor nanoparticles. A 3-nm particle emits green light at 520 nm, whereas a slightly larger particle, measuring 5.5 nm in diameter, will emit in the red, around 630 nm. Courtesy of D.J. Norris et al.

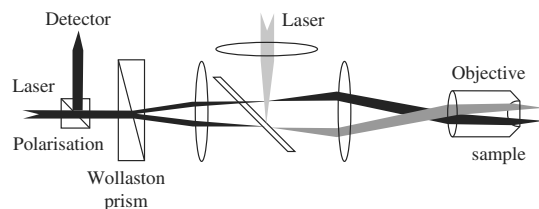


**Fig. 17.14.** Grafting a biomolecule onto a biological nanoparticle

hydrocarbons, etc., substituted at the silicon site. Such substitutions constitute a first step towards further biochemical conversion. Silanisation lends itself to the addition of phosphonate, polyethylene glycol, or ammonium groups whose presence on the nanoparticle surface ensures its water solubility. Including phospholipid nanoparticles equipped with long hydrophobic tails leads to a double-layer segregation, the nanoparticle being enclosed within the inner cavity and the outer face then rendered hydrophilic. In both cases, the fluorescence yield remains unchanged, but with the advantage of aqueous solubility and long-term stability.

The cross-linking element is part of the kit already discussed in more detail with regard to the linking of biological structures (e.g., biotin, DNA strands, proteins, etc.). Such structures are then accessible to the FISH method already mentioned and their use has witnessed a tremendous upsurge recently.

Finally, let us note the importance of metallic nanostructures. These make up another large emerging family, complementary to semiconductors and molecules. Gold is the most widely used metal, followed by silver, although its surface state is more difficult to stabilise. Electronic resonances of surface



**Fig. 17.15.** Photothermal detection of metallic nanoparticles

plasmon type (see Chap. 16) are characteristic of interfaces between the metal and a dielectric structure such as a biological medium. By grafting a metallic nanostructure such as gold onto a target oligonucleotide, one can considerably reduce the optical detection threshold for pairing of complementary strands, one of which is attached to a solid substrate.

More recent and very promising work involves photothermal detection of local heating photoinduced by absorption of light on a gold nanostructure. Figure 17.15 illustrates the detection principle based on interferometry. The temperature increase due to illumination by the 514-nm line of an argon laser (injecting a continuous power of the order of 20 mW into the material) is probed by a He–Ne laser at 633 nm. Nanoparticles with sizes as small as 2 nm can thereby be imaged.

The resulting temperature increase, estimated at a few kelvins for 5-nm particles, is doubtless still a little too high to be acceptable without damage in a biological medium. However, improvement in detection sensitivity should make it possible to reduce this level significantly. This highly promising method avoids the drawbacks intrinsic to organic structures (tendency to photobleaching), but also those that create difficulties when using semiconductor nanostructures (intermittent emission, known as blinking).

### 17.2.3 Molecular Engineering for Biophotonics

The high number of degrees of freedom displayed by molecules make them eminently suitable for functionalising nano- and microscopic constructions composed of tailor-made molecules, i.e., molecules fulfilling a detailed set of physicochemical specifications making them compatible for application to a biological environment. For this purpose, a predictive relation must be established between the nature of these molecules, but also their orientation and relative disposition in space, and the existence of the effect or combination of effects one seeks to implement. And in addition to this, the intensity or yield of these effects must be optimised, or else the initiation or observation threshold must be reduced.

These constructions are capable of scanning all length scales, from the molecule to the bulk material, passing through the nanocluster and the microstructure. Furthermore, they are of all different kinds:

- Purely organic, as happens when the relevant entity, e.g., an aggregate, is composed of standard dyes such as rhodamine or fluorescein.
- Organometallic. Ruthenium–trisbipyridine complexes play an important role in biophotonics through their affinity with certain specific sites on DNA and their photoreactivity. These properties make them candidates for use in genetic phototherapy.
- Organomineral. The biomineral crystallisation of organic entities such as proteins with inorganic crystals such as calcite or apatite is omnipresent in cartilaginous tissue such as tooth enamel or the bony part of the skeleton.
- Amorphous, non-crystalline media, such as polymers, e.g., polystyrene or poly(methylmethacrylate) (PMMA), possibly in the form of nanospheres within which functional entities with the required photonic properties are dispersed or attached.

In the following, we shall be concerned with the link between the structure of molecules and the optimisation of various optical properties arising from this structure, without considering at this stage the effects of interactions between molecules or of their spatial arrangement. Indeed, the intermolecular binding energy can be neglected to first order compared with the dominating effect of intramolecular chemical bonds which are more cohesive and energetic. Spatial arrangement raises some delicate and novel problems that are characteristic of the nanoscale and less important or even absent on longer length scales. Consider the case of a structure whose characteristic dimension is commensurable with a significant variation in the electromagnetic field interacting with it, i.e., it has characteristic dimensions in the range from  $\lambda/10$  to  $\lambda/20$ . Here the constraint of non-centrosymmetry which applies so clearly in large scale bulk materials with a view to second harmonic generation loses all validity. Indeed, other types of coupling, such as quadrupole coupling of the structure with the electric field gradient, or surface effects of the same order of magnitude as bulk effects on this length scale, are likely to efficiently make up for the absence of centrosymmetry breaking and to induce the same effects.

In this section, we consider three major categories of molecular systems whose generic structures, illustrated by several specific examples in each case, are shown in Figs. 17.16 and 17.17:

- Nonlinear molecules. We consider only quadratic effects, the lowest order nonlinear effect. These molecules can be divided into two basic families: dipolar and octupolar molecules, from which multipolar systems can be generated. They have just arrived on the nanobiophotonic stage but prospects for them are good, especially in the mapping of membrane potentials or potentials in the vicinity of neurones.
- Molecules suitable for two-photon absorption, i.e., displaying a certain efficiency in the ‘simultaneous’ capture of photon pairs by absorption (in the quantum mechanical sense, as predicted in the 1930s by Maria Goeppert-Mayer), with the return to the ground state occurring by the more classical

channel of fluorescent emission. These molecules are becoming more common in conjunction with the development of confocal microscopy in biology labs (see Sect. 17.3). The luminescence generated by the pump laser comes mainly from an elementary region surrounding the focal point, defined from the diffraction properties of the beam, either Gaussian or otherwise, and the definition of the focusing system. This property, a consequence of the quadratic dependence of two-photon absorption, and hence of the emitted fluorescence, as a function of the incident beam intensity, has the benefit of eliminating any significant background contribution from autofluorescence by tissues along the optical path on either side of the elementary volume (as would have been the case with one-photon fluorescence). The contrast and readability of images is thereby greatly enhanced.

- The fluoroionophores, a family of molecules whose luminescence properties are controlled by complexation to a specific target entity such as an ion, and which can thus be incorporated into the composition of new types of ultrasensitive and specific sensor, e.g., nanoscopic pH meters or calcium ion detectors, recalling that calcium, potassium or sodium flows and their variations are key features in the life of a cell.

The first category of molecules here, i.e., those with strong quadratic nonlinearity, has been under constant scrutiny for over three decades now. The initial motivation for these studies came from optical data processing for telecommunications and in particular, electro-optical modulation and switching, which together with microlasers will play a key part in very high speed data transmission systems. Indeed, these aims are more than ever relevant today in view of the issues at stake and the constant shift of objectives towards improved performance, particularly with regard to broadband issues. The first applications to cell imaging are beginning to appear, affording a new and extensive field of application to these systems, alongside complementary techniques that are already commonly used, such as one- and two-photon fluorescence.

The cubic anharmonicity of the electron polarisation under the effect of the electric field of a laser beam generates in its turn radiation at all frequencies that are multiples of the fundamental frequency, i.e.,  $2\omega, 3\omega, \dots, n\omega$ . The most intense radiation corresponds to the harmonic of lowest order, i.e., the doubled frequency (provided that the symmetry of the system allows this emission). The (cubic) anharmonic potential felt by the electrons during their excursions under the effect of the exciting field results from strong centrosymmetry breaking brought about by jointly attaching a donor group conjugate to an acceptor group through a centrally located system made up of highly polarisable delocalised electrons. The central part of the molecular system, sometimes called the  $\pi$  transmitter (see the diagram of the generic structure in Fig. 17.16), is generally provided by aromatic or polyaromatic systems. These are well suited to the requirement that the electrons should be highly delocalised, making them sensitive, through their displacements and subsequent spatial explorations beyond the central quasilinear part of the internal

potential, to anharmonic distortions of the surrounding potential. This anharmonicity shows up in turn through the radiation induced by these electronic displacements in the form of even harmonics including  $2\omega$ . In this context, one may legitimately speak of molecular antennas and the optimisation of their anharmonicity.

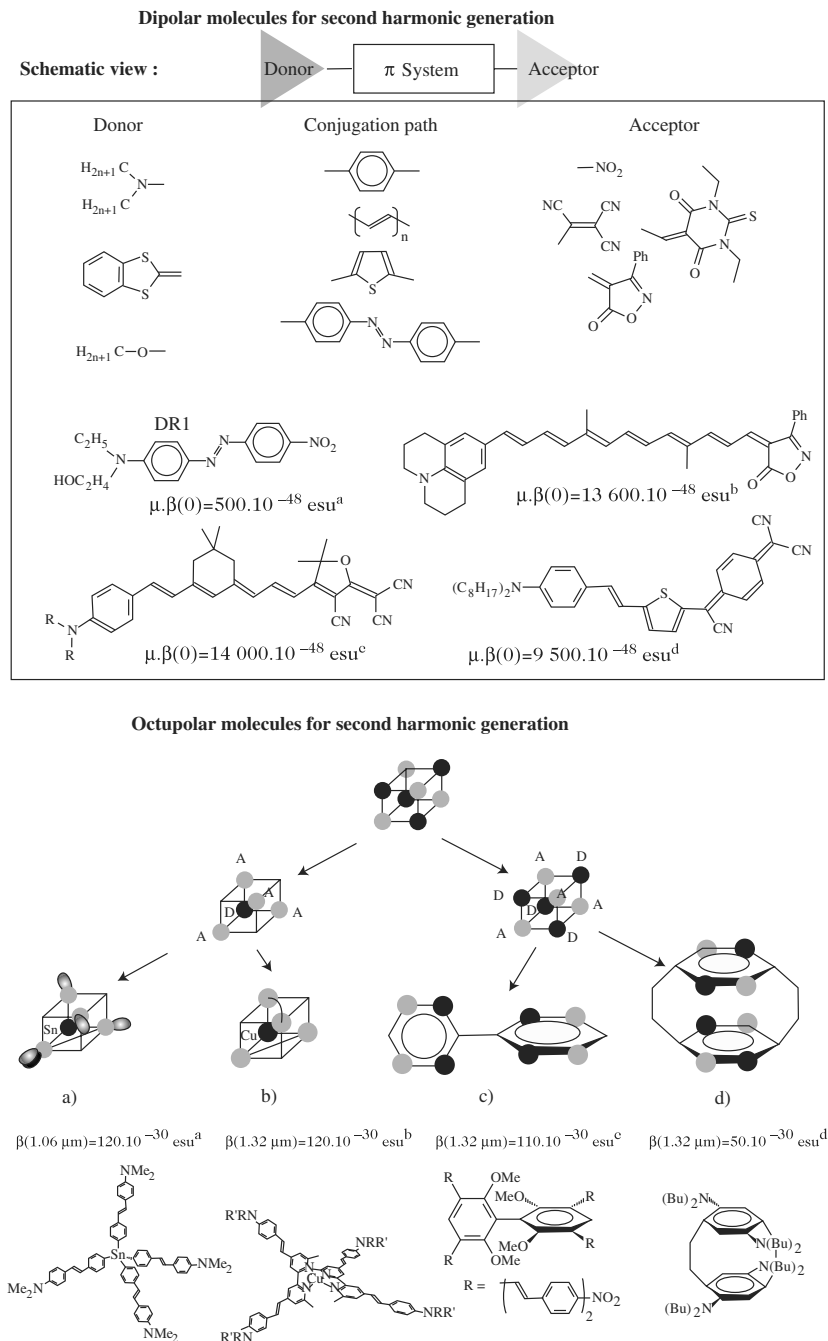
A two-level quantum model was proposed in the 1970s and has been continually improved since then to account for new structures. Hence, it was shown that a 3-level model is required to describe 2D octupolar systems, and a 5-level model to generalise this to 3D. To a first approximation, the two-level model can be used to account for the anharmonicity of the induced anharmonic polarisation, leading to the following expression for the coefficient  $\beta$  (strictly speaking, a tensor) relating the harmonic polarisation to the square of the fundamental field:

$$\beta = \frac{\mu^{2\omega}}{E_\omega^2} = \frac{3\hbar^2}{2m} \frac{\mu_{01}^2 \Delta\mu}{[\hbar^2\Omega^2 - (2\hbar\omega)^2](\hbar^2\Omega^2 - \hbar^2\omega)}, \quad (17.19)$$

where  $\mu^{2\omega}$  is the amplitude of the electron dipole induced at frequency  $2\omega$ , and  $E_\omega$  is the amplitude of the fundamental field. The terms of (17.12) can be recognised here. The excited level is denoted here by  $|1\rangle$ , whilst  $\Delta\mu$  is the difference in dipole moment between the ground and excited states.  $\mu_{01}$  is the transition dipole between the ground state and the excited state,  $\hbar\Omega$  is the energy separating the ground state and the excited state (often called the gap), and  $\hbar\omega$  and  $2\hbar\omega$  are the energies of the fundamental and harmonic photons, respectively.

Referring to the generic model illustrated in Fig. 17.16, a certain number of molecules materialise particular combinations of the three basic components of the system, and experimental values of the parameter  $\mu\beta$  are given. The combination of the dipole moment  $\mu$  of the ground state with  $\beta$  reflects the suitability of these entities to orientate themselves in an electric field in the framework of the Langevin model as implemented in the context of the EFISH (electric-field-induced second harmonic) method. This allows measurement of the parameter  $\beta$  by centrosymmetry breaking of an ensemble of molecules in solution at thermodynamic equilibrium resulting from the dipole interaction with an electric field via the permanent ground state molecular dipole.

The rod-shaped bipolar structure of donor–acceptor type is in fact just a particular limiting case of more general multipolar structures, whose other limiting case is associated with the so-called octupolar structures. The indices  $J = 1$  and  $J = 3$  are a conventional notation arising historically from symmetry considerations in quantum mechanics as applied to atomic physics (addition of angular momenta and pioneering work by E. Wigner, who applied group theory to physics as early as the 1930s). Indeed, one may decompose the tensor  $\beta = \beta_{J=1} + \beta_{J=3}$  into its two independent components  $\beta_{J=1}$  and  $\beta_{J=3}$ , also called its irreducible components. The first can be associated with a spatial average of the charge density of dipolar character, corresponding to the operator  $x$ , which, all things being equal, amounts to assimilating a

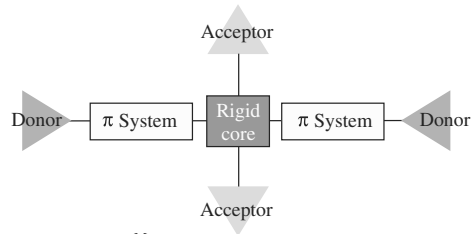


**Fig. 17.16.** Optimisation of molecular structures for generation of harmonics. From [24]

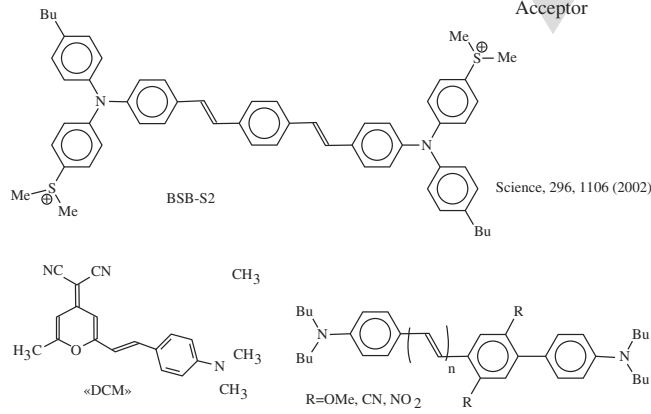


Molecules optimised for two-photon absorption

Schematic view :  
Centrosymmetric quadrupolar structure

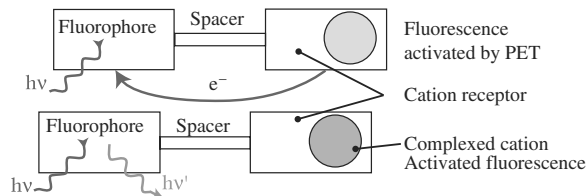


Examples :



Fluorescent molecules for detecting ions by transfer of photoinduced electrons versus fluorescence

Schematic view :



Exemples :

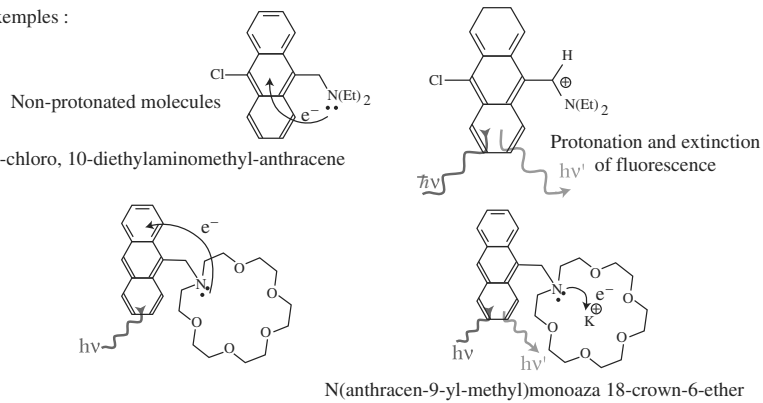


Fig. 17.17. Optimisation of molecular structures for two-photon absorption and photoinduced effects. From [24]

pear-shaped charged object to a rod, but weighting the swollen end more heavily. The other component is associated with a deviation from this dipole approximation, which can be shown to correspond to an octupolar structure, i.e., a system with order 3 symmetry, associated with an average over the charge density of the cubic operator  $x^3$ . The most general model for an octupolar system can be assimilated with a cube whose eight corners are alternately occupied by donor and acceptor groups identical to those already mentioned for the dipole structures, i.e., four electron donor groups alternating with four electron acceptor groups (see Fig. 17.17).

Such structures have considerably extended the previous field and raise questions that are both fundamental and practical, concerning the absence of a dipole characterising octupolar structures. One of the main issues concerns the impossibility of macroscopic centrosymmetry breaking by orientation in an electric field, owing to the intrinsic absence of a dipole that might allow such coupling. The answers to these questions, which belong to the burgeoning field of quantum coherent control applied to the 'all-optical' manipulation of molecular systems, goes beyond the scope of this introductory discussion. The interested reader is referred to the research literature or more specialised texts.

These various challenges have proved extremely fertile and have generated a new interface between nonlinear optics, a major area of photonics, and biology. Indeed, previous methods for orientating molecules in solution (the Langevin method associated with the coupling between the orientating electric field and the permanent dipole) excluded the ionic species that are omnipresent in biology, making it very difficult, if not impossible, to operate in an aqueous medium, whilst imposing very restrictive experimental configurations. Indeed, a capacitive cell is used to generate an impulsive orientating field with pulse widths of microsecond order and weak periodicity of the order of a few cycles or tens of cycles per second to avoid the risk of hydrolysing or decomposing the solution (even if neutral).

The alternative approach, now widely adopted, is the harmonic scattering of light, also called hyper-Rayleigh scattering. The incoherent character of this light provides a way of overcoming the constraint of breaking the centrosymmetry of the medium, but the signal is much weaker and requires careful optimisation of the detectors. Indeed, whereas the harmonic signal generated by the EFISH techniques displays a coherent character, hence proportional to the square of the concentration of species in solution, the incoherent nature of the signal produced by anharmonic Rayleigh scattering leads to a weaker response, since it is simply linearly proportional to this concentration of active species. The possibility of detecting this response and following it over time (nonlinear monitoring) in a medium that may be aqueous or acidic as happens in biologically relevant situations, whether one is dealing with ionic species with or without dipoles, which may evolve over time (e.g., complexation of species, protonation or deprotonation which can directly and sensitively influence the donor-acceptor charge transfer underlying  $\beta$  and thereby open an

invaluable observational window upon the cellular medium), has led to a new field of investigation in the area of nonlinear phenomena in biological media.

The tensors  $\beta$  and their spectral dependence in different wavelength ranges depending on the application are tending to become fundamental data for qualifying chemically and biochemically relevant molecules, complementing the information inferred from more conventional constants such as optical rotating power, dielectric constant, molar extinction coefficient, and their spectral dispersions. The cubic nonlinear effects associated with the perturbation of the dipole of order immediately greater than  $\beta$  and corresponding to a quartic perturbation of the harmonic potential will not be discussed here. They are responsible for important effects, such as third harmonic generation or the effects known as four-wave mixing. The reader is referred to the research papers indicated in the bibliography.

The second family of molecules illustrated in Fig. 17.17 is the family of fluorescent molecules with high quantum yield for two-photon absorption, the principles of which have already been described. The optimisation of two-photon absorption is the central part of the problem here, more critical than optimising the yield of the subsequent fluorescence. Organic dyes generally exhibit good fluorescence yields, sometimes close to unity. Such a level corresponds to the theoretical optimal situation of one fluorescence photon per absorbed pump photon, and in no way implies that the measured effective yield will be 100%, given the rather unpredictable geometric path of the photons, which may lead to inefficient collection of fluorescence photons on the sensitive surface of the photodetector, the limited efficiency of the detectors themselves, and even the limited absorption cross-section of the fluorescent molecule with regard to the pump photons. However, the picture here may be less demanding than for the optimisation of quadratic properties, where the lack of centrosymmetry is a fundamental constraint on the molecular level.

Systematic studies have brought out a generic structure of quadrupole type. From consideration of tensor symmetries, it can be established that the multipole components of even order, e.g., charge  $J = 0$ , quadrupole  $J = 2$ , hexadecapole  $J = 4$ , are associated with different molecular geometries relative to the optimisation of the tensor  $\gamma$  (imaginary part), the imaginary part of which is associated with two-photon transitions. A model linear structure comprises a conjugated chain with two identical electron donor groups attached symmetrically at its ends, whose action can be reinforced by two other acceptor groups attached to a rigid, central core, e.g., an aromatic cycle, located at the centre of the conjugated system, producing a quadrupole-type structure.

The two examples in Fig. 17.17 show ways of making fluorescent systems according to this scheme. The two-photon absorption cross-section can reach  $10^{-47}$  cm<sup>4</sup>s per photon for the factor  $\delta$  given by

$$\delta = \frac{3\hbar\omega^2 \text{Im}\gamma_{\text{NLI}}}{\varepsilon_0 c^2}, \quad (17.20)$$

where the cubic polarisability  $\gamma_{\text{NLI}}$  corresponding to the index depending on the incident beam intensity was given in (17.2).  $\delta$  is of the order of  $10^{-50} \text{ cm}^4\text{s}$  per photon for standard non-optimised two-photon markers still in use a few years ago.

A particularly interesting case in nanotechnology is provided by the second molecule BSB-S<sub>2</sub> which has an extra rather special functionality. Indeed, the two-photon absorption phenomenon, particularly efficient for this molecular system, is followed by emission of protons which can then themselves initiate chemical reactions in the irradiated medium, following a sudden and highly localised pH variation. Such molecules are known as two-photon acid photogenerators (APG). They have been used, for example, to initiate a photopolymerisation reaction locally at the focal point of a confocal microscopy device, this reaction involving only the submicron focal region of the device (see Sect. 17.3). Combined with the advantage that the threshold for the photopolymerisation effect is higher for two-photon absorption than for normal one-photon absorption, it becomes possible to achieve truly nanometric writing in the highly irradiated focal region of photosensitive polymers such as acrylates or epoxides within which APG-type molecules have been previously dispersed, capable of efficiently absorbing photon pairs.

Hence, with the molecule BSB-S<sub>2</sub> and other APG systems, the fluorescence yield collapses in favour of the highly efficient release of a proton by the lateral dimethylsulfonium groups under two-photon excitation. The protons photoemitted in this way will then initiate polymerisation of the epoxides and eventually lead to exquisite photoencoded nanostructures.

Finally, let us mention the special case of the standard dye molecule known as DCM [4-dicyanomethylene-2-methyl-6-(*p*-(dimethylamino)styryl)-4H-pyran], which is important in practice. It combines the benefits of a high- $\beta$  nonlinear structure ( $\beta \approx 20 \times 10^{-30} \text{ esu}$  at  $1.064 \mu\text{m}$ ), due to a pair of donor-acceptor groups connected through a linked system of  $\pi$  electrons, while displaying significant one- and two-photon fluorescence quantum yields. This is a phototypical example of a multifunctional system with the essential practical advantage that it is robust under irradiation by a pump laser beam in the visible. Indeed, it has long been used in dye lasers and is commercially available. This kind of multifunctionality has become relevant again in molecular engineering or optical tracking of molecular traffic (the colourful term adopted in cell biology) in biological media via the complementary triple signature of second (or third) harmonic generation and one- and two-photon fluorescence.

The third illustrative family mentioned earlier comprises molecules for the detection of ions or other species by transfer of photoinduced electrons, which tends to inhibit fluorescence effects. Their generic structure, shown in the lower part of Fig. 17.17, has three modular parts:

- a fluorescent subentity that is active in the absence of the ion target,
- a spacer of variable effective length and type,

- a cryptand subentity designed to display a strong selective affinity with a certain type of species to be dosed (in this case, cations, e.g., protons).

When this complexation occurs, the electron from the receptor which otherwise tended to deactivate fluorescence is preferentially captured by the cation and the return to the ground state by emission of a fluorescence photon is now authorised.

Two examples are shown in the figure. The first illustrates protonation of an anthracene derivative (a tertiary amine playing the role of proton receptor), allowing very sensitive pH measurement in biologically or chemically relevant environments by the method described above. The second example illustrates the more sophisticated chemistry of the crown ethers (due to Cram et al. and perfectly representative of the generic family of the cryptates). The example shown complexes the potassium ion, which is particularly important in biology, but it could equally well be the divalent calcium cation in another configuration.

In each example, the free electron pair of the nitrogen, which occurs in a so-called non-binding molecular orbital (this term refers to the high lability of the electrons attached to it, located energywise between the two extreme cases of filled bonding orbitals and empty antibonding orbitals), are able to orient themselves either towards the cation, thereby stabilising it, or in its absence, towards the fluorophore part of the molecule, thereby opening up a non-radiative deexcitation channel for the system.

### 17.3 Nanophotonic Instrumentation for Biology

In order to improve our understanding of the more complex biological mechanisms, one must find out how molecules interact on their own length scale. Recent experiments carried out on the scale of single molecules have brought us closer to such processes. They are based upon engineering, detection and manipulation of molecular and macromolecular entities or particles in typical media, e.g., living cells or artificial membranes. In the present section, we shall describe various techniques that can be used to measure the properties of single molecules in biology, i.e., optical and mechanical properties. The advantage in detecting isolated entities is the access gained to heterogeneous behaviour that would be hidden within an averaging measurement made over an ensemble, such as specific fluctuations due to the interaction of a molecule with its immediate surroundings, for example. The following sections outline conditions for observing single molecules by fluorescence and applications connected with this technique.

#### 17.3.1 Optical Detection of Single Molecules by Fluorescence

The advantage with optical measurements is that there is no mechanical contact, so that molecules can be studied from a distance and without

perturbations that would be induced by a tip, for instance. Techniques for collection and detection of very weak optical signals have progressed enormously, and today it is possible to detect isolated molecules by fluorescence or Raman spectroscopy. Historically, the first single molecules were observed by fluorescence at very low temperature in crystals. This opened up new prospects in molecular spectroscopy, quantum optics and condensed matter physics [25–29]. Many studies at room temperature were to follow, with a very wide field of applications, since it now became possible to image isolated proteins in cells, to observe the dynamics of excitation transfer between two entities, and to measure the lifetimes of single molecules [30–33].

### Basic Principles and Techniques for Detecting Single Molecules by Fluorescence

#### *Detecting and Imaging Single Molecules*

The main limitation when detecting a single molecule is its very weak fluorescence signal, which must be isolated from all other signals arising in the neighbourhood. One must therefore ensure that only one molecule is detectable, that it is in resonance with the exciting laser, and that its fluorescence emission is efficient. In a typical single-molecule detection experiment, a laser beam with excitation power  $P$  and energy  $h\nu$  is focused on the sample through a first optical element, an objective in microscopy or the tip of a fibre in near-field optics. Focusing occurs over an area  $A$  that is determined by the size of the diffraction pattern in the case of illumination through an optical element ( $A$  may be more difficult to assess for illumination under a tip or by evanescent waves).

The total efficiency of the excitation–emission–detection process for a single molecule depends on two main parameters:

- The absorption cross-section  $\sigma$  of the molecule. This is defined as the area over which it can ‘absorb’ an incident beam. The quantity  $\sigma$  can be rigorously related to (17.6) in Sect. 17.1 applied to an isolated molecule.
- The fluorescence quantum yield  $Q$ . This is the number of emitted photons per absorbed photon.

The signal emitted by a molecule is then

$$\langle F(t) \rangle = CQ \frac{\langle P(t) \rangle \sigma}{Ah\nu} \quad [\text{counts (or photons) per second}], \quad (17.21)$$

where  $C$  is the collection efficiency of the optical system,  $h\nu$  the energy of an incident photon, and  $\langle P(t) \rangle$  the mean excitation power over the detection time. One can also define the intensity  $\langle I_0(t) \rangle = \langle P(t) \rangle / A$  in  $\text{W}/\text{cm}^2$ , where  $A$  is the focal area on the molecule.

Superposed on this signal is the background noise from the surroundings or the substrate: residual fluorescence emission if any, Raman or Rayleigh

scattering, and electronic noise in the detector. Moreover, to be certain that only one molecule has been excited, one must ensure that it is sufficiently far away from the others in the sample, by dilution, or one must reduce the area of focusing (or excitation in the optical near field). The various optical techniques used for illumination are outlined in the next section.

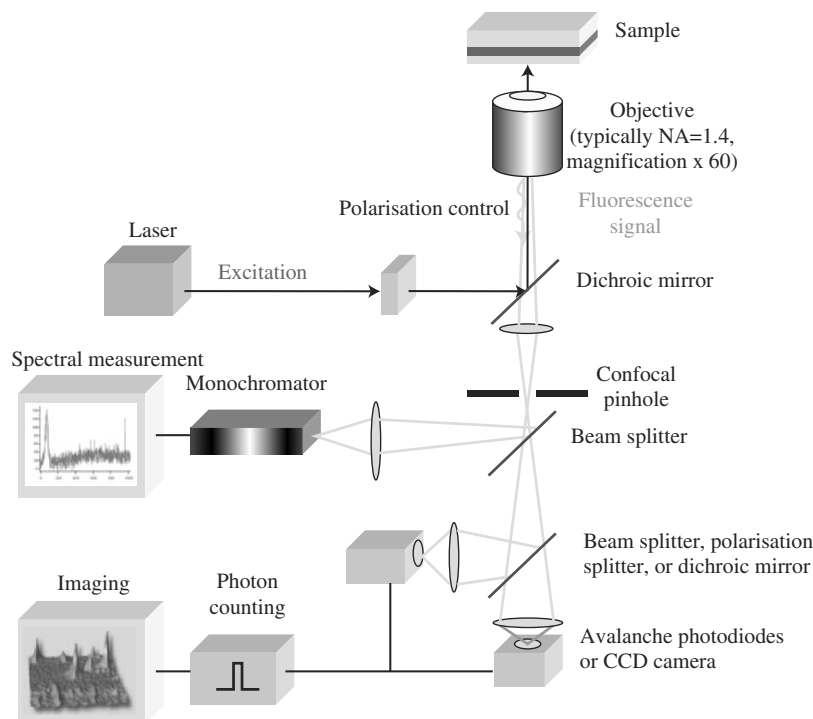
Typically,  $\sigma = 2.8 \times 10^{-16} \text{ cm}^2$  (comparable with molecular sizes) and  $Q = 0.85$ , for a fluorescein isothiocyanate (FITC) molecule excited at  $\lambda = 525 \text{ nm}$  and diffusing freely in water. For comparison, the Raman effect of the molecular environment can reach a cross-section of the order of  $10^{-12} \text{ \AA}^2$  for benzene, for example. With confocal detection (see Fig. 17.16), a typical value of the collection efficiency  $C$  is 2–10%. (Naturally, only a part of the fluorescence emitted in all space directions is actually intercepted.) Excitation areas can vary. In confocal detection, a typical value is  $A = 1 \mu\text{m}^2$  (diffraction limit in a microscope using a wide-aperture objective). In near-field microscopy,  $A$  can be as low as  $0.01 \mu\text{m}^2$  (see Fig. 17.16). Using an excitation power of 100 mW and an integration time of 1 ms in confocal microscopy, one can thus obtain signals of 66 photon/ms.

Noise from the environment is typically 100 Hz/ $\mu\text{W}$ , or equivalently 100 photon/s/ $\mu\text{W}$ , whilst the electronic noise in sensitive detectors like avalanche photodiodes is of the order of 100 Hz. The overall noise level thus corresponds to 7 photon/s in the conditions described above, which remains well below the expected signal from an isolated fluorescing molecule.

The detectors commonly used in single-molecule spectroscopy operate in photon counting mode. Avalanche photodiodes are often used. They have the advantage of a high quantum yield (around 70% in the visible) and a time resolution of the order of 400 ps. Photomultipliers have better time resolution (around 20 ps), but lower quantum yield (10–20%).

Despite their reasonable emission rates, the detection of single molecules is limited by their finite emission time. In the aqueous environments used in biology, which are therefore rich in oxygen, the fluorophores can form non-fluorescent radicals by photochemical reactions in the excited state with other reactive molecules such as oxygen. These reactions irreversibly alter the chemical nature of the molecule, leading to a sudden interruption in the fluorescence emission. Before this interruption, known as photobleaching (see Sect. 17.1), an FITC molecule in water can typically emit  $10^6$  photons, reducing to 15 s the possible observation time for FITC molecules that have been immobilised for study in the conditions described earlier. One of the main aims in molecular engineering of fluorescent molecules is to enhance their stability.

Prior to this sudden disappearance of the signal, the time evolution of the emission from a single molecule reveals the deexcitation cycles that it undergoes: the emission is sometimes interrupted when it passes through non-radiative states, e.g., the triplet state, and this leads to an intermittent fluorescence signal. These non-radiative periods degrade the fluorescence quantum yield. Indeed, if the molecule underwent no transitions into these non-radiative states, the rate of photon emission could easily be deduced from



**Fig. 17.18.** Setup for fluorescence detection of single molecules. Measurement of emission spectrum, imaging, polarisation resolution (a polarisation splitter separates the signal between the two detectors), excitation transfer resolution (a dichroic mirror separates two wavelengths detected by each photodiode)

the fluorescence lifetime  $\tau_S$ , where by definition

$$\tau_S = \frac{1}{k_r^S + k_{nr}^S},$$

with the notation of Sect. 17.2. A typical value of the fluorescence lifetime of a molecule is  $\tau_S \approx 5$  ns, which gives a fluorescence emission rate of about  $10^8$  photon/s. With a rather modest collection efficiency, it would thus be possible to detect around  $10^6$  photon/s. Unfortunately, this fluorescence rate is often limited by access to non-radiative states during decay. The molecules are also sensitive to their environment, and time variations in the fluorescence emission of signal molecules are the signature of these interactions with their immediate environment. These properties are used to provide information about molecular behaviour in complex media.

#### *Microscopy Techniques for Detecting Single Molecules*

Microscopy techniques have been improved considerably over the last few years, especially with the advent of high resolution microscopes using objectives of excellent optical quality and large numerical aperture, but also



thanks to near-field techniques. Fluorescence detection is now possible in biological samples with a lateral resolution of a few hundred nanometers. The longitudinal resolution has also been improved using techniques such as confocal microscopy or two-photon microscopy (see Sect. 17.3.2). Different instrumental setups for microscopy based on one-photon fluorescence are shown in Figs. 17.18–17.21. Most fluorescence microscopy techniques work by reflection. One then speaks of inverted microscopy, in which the optical image of the sample is often found point by point, in contrast to what happens in traditional fluorescence microscopy by parallel imaging. A 2D image of the sample is obtained either by sweeping the sample across the fixed focal point of an objective using a piezoelectric setup, or by scanning the laser beam across the fixed sample by means of mirrors mounted on a galvanometric system. In inverted microscopy, the objective used to focus the beam on the sample also serves to collect the signal, thereby significantly simplifying the optical setup.

Figure 17.18 shows a general optical setup for detecting single molecules. More specific excitation geometries are discussed below.

### Experimental Techniques for Spatial Selection of a Single Molecule

In a typical single-molecule detection experiment, a laser beam with excitation power  $P$  and frequency  $\nu$  is focused on the sample through a first optical element, e.g., the objective for a microscope, or the tip of a fibre in the optical near-field case. Focusing occurs over an area  $A$  determined by the size of the diffraction pattern. The signal is collected by a second optical element which may be the same as the one used for excitation (see Fig. 17.19).

For excitation by the optical near field, the laser beam passes through a tapered fibre (see Fig. 17.20, left). At the end of the fibre, it emerges in the form of an evanescent wave whose intensity falls off exponentially over several tens of nanometers. The excitation area depends sensitively on the shape of the fibre tip and its distance from the sample.

For excitation under total internal reflection (see Fig. 17.20, right), the angle of incidence of the beam on the sample is greater than the critical angle of grazing

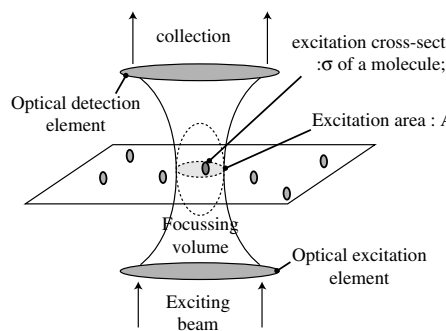


Fig. 17.19. Excitation with large numerical aperture

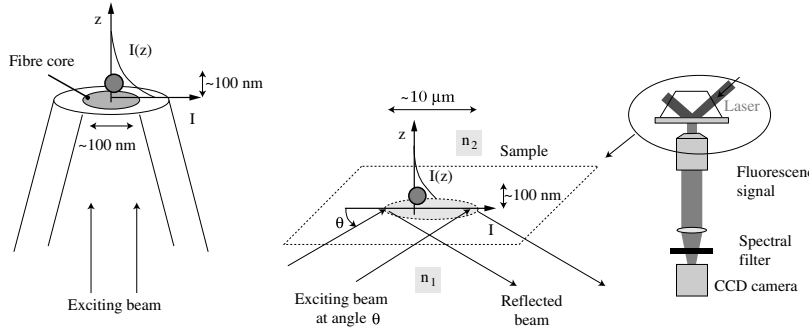


Fig. 17.20. *Left*: Near-field excitation. *Right*: Excitation by total internal reflection

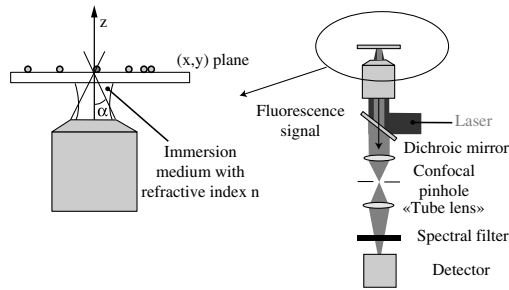


Fig. 17.21. Excitation in the confocal configuration

refraction between the first medium (glass of the substrate) with index  $n_1$  and the medium containing the molecule (liquid, air, etc.) with index  $n_2$ . This angle  $i_m$  is defined by the condition  $n_1 \sin i_m = n_2$ . There is then an exponential decrease of the incident intensity in the medium over a wavelength  $d$  such that

$$d = \frac{\lambda}{4\pi} \sqrt{n_1^2 \sin^2 \theta - n_2^2} \approx \frac{\lambda}{5} .$$

The quantity  $NA = n \sin \alpha$  is the numerical aperture of the objective or focusing lens. In confocal microscopy (see Fig. 17.21), a confocal diaphragm in the image plane improves longitudinal resolution by removing light rays arising from emission from other planes than this object plane. The lateral diameter of the focusing spot is of the order of  $0.51\lambda/NA \approx \lambda/2$ . In the longitudinal direction, the size of the focusing spot is less than  $n\lambda/NA^2$ .

*Confocal Microscopy.* The distinguishing feature of the confocal microscope is the confocal pinhole, which serves as detection aperture and is optically conjugate to the sample point under observation (see Figs. 17.21 and 17.22). The confocal microscope can thus image samples in three dimensions by rejecting the fluorescence background arising from the environment of the fluorescent entities that interest us, but also excluding photons scattered from the focal point. Ideally, the confocal pinhole should be similar in size to the diameter of the diffraction pattern arising from the image of a point object by the

objective. Hence, the main drawback with this technique is the loss of light intensity due to the diaphragm.

The resolution of a microscope must be defined according to some specified criterion. One can choose either the size of the image of an infinitely small point source (full width at half maximum or FWHM, which is the diameter of the light spot where the intensity descends to half of its maximum value), or the minimum distance between two point sources at which they can still be distinguished (Rayleigh criterion), or the passband in spatial frequencies for an optical instrument (a notion arising from Fourier optics). In the case of the Rayleigh criterion, two point sources can first be distinguished when the maximum of the image of one coincides with the first minimum in the diffraction pattern produced by the other. The spatial intensity distribution near the focal point for an initially homogeneous distribution of the amplitude front before focusing is given by the function  $I(u, v)$ , where  $u$  and  $v$  are the reduced spatial coordinates defined below. The analytical solution is

$$I(0, v) \propto \left| \frac{2J_1(v)}{v} \right|^2 \quad (\text{Airy disk}), \quad (17.22)$$

where  $J_1(v)$  is the Bessel function of order 1,  $v = rn2\pi \sin \alpha/\lambda$  is the reduced radial coordinate with  $r$  the radial coordinate,  $n$  is the refractive index of the medium in which focusing occurs, e.g., a refractive index oil,  $\lambda$  is the incident wavelength in the vacuum, and  $\alpha$  is the half-angle of the aperture cone of the objective, defined above.

Similarly, the intensity distribution along the optical axis is given by

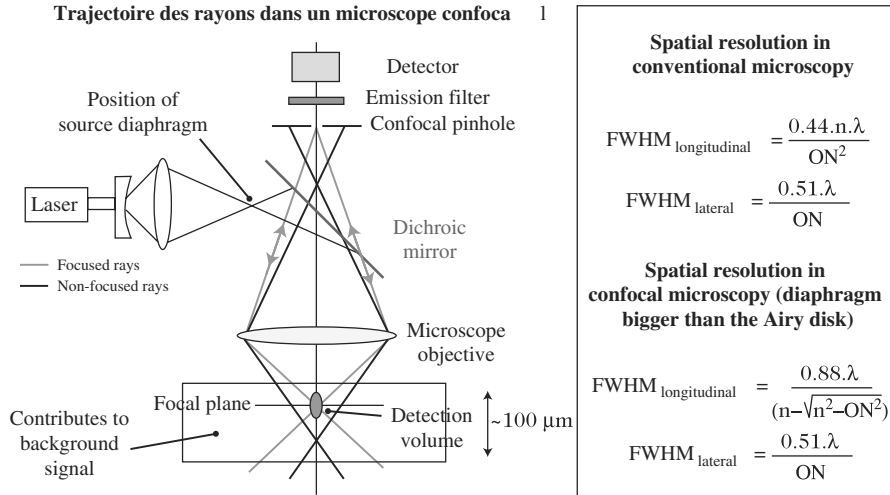
$$I(u, 0) \propto \left[ \frac{\sin(u/4)}{u/4} \right]^2, \quad (17.23)$$

where  $u = zn2\pi \sin^2 \alpha/\lambda$  is the reduced longitudinal coordinate and  $z$  is the longitudinal coordinate.

The first zero of the function  $I(u, v)$  is located at  $(u_0 = 4\pi, v_0 = 1.22\pi)$ , and the lateral and longitudinal FWHM values of the spot are given in Fig. 17.22.

It is clearly better to work with larger numerical apertures. This is why high index immersion oils are often used between the objective and the sample (typically  $n \approx 1.5$ ). For  $\text{NA} = 1.4$ ,  $\lambda = 500$  nm, one obtains lateral FWHM = 440 nm and longitudinal FWHM = 765 nm. Note also that the expressions given here arise from the plane wave diffraction theory, which is not strictly applicable for large numerical apertures. The full diffraction theory predicts slightly smaller values than the classical theory.

The true resolution of a microscope is not purely defined by the effective size of the interaction volume. It also depends on the optical configuration used for detection and the coherence of the signal. The global point spread function (PSF) results from the optical elements and diaphragms under illumination  $\text{PSF}_{\text{ill}}(r, z)$  and during detection  $\text{PSF}_{\text{det}}(r, z)$ . Then, if a source producing a



**Fig. 17.22.** Schematic of the light rays propagating through a confocal microscope. The objective focuses light from a laser onto the sample. Fluorescence from the sample is collected by the same objective and detected by a photodetector, e.g., a photomultiplier or an avalanche photodiode. A dichroic mirror then transmits the fluorescence since it reflects wavelengths shorter than the excitation wavelength. By virtue of the confocal pinhole, only those rays arising exactly from the image focal plane are actually collected. FWHM = full width at half maximum. From [34]

spatial light distribution  $O(r, z)$  is imaged, the final image will have intensity distribution

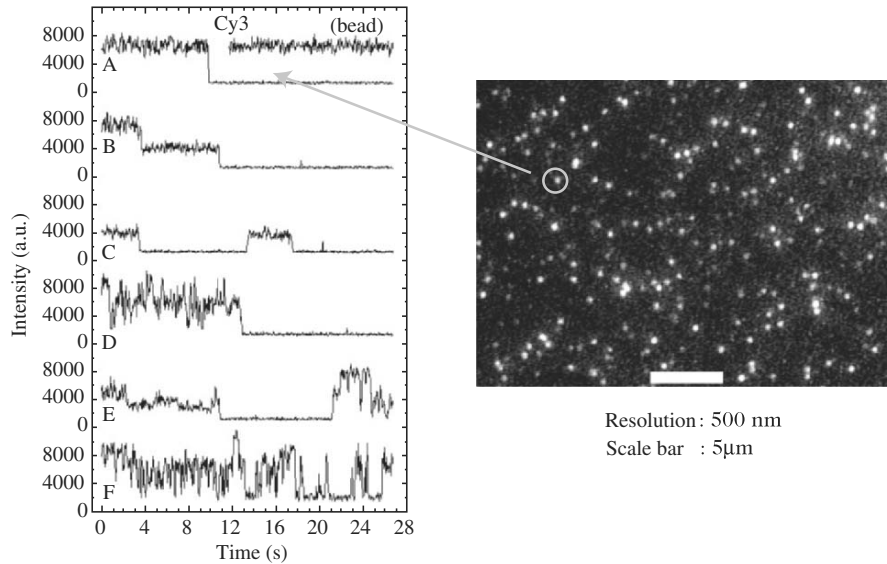
$$I(r, z) = \left[ \text{PSF}_{\text{ill}}(r, z) \text{PSF}_{\text{det}}(r, z) \right]^2 \otimes O^2(r, z), \quad (17.24)$$

for an incoherent optical process, e.g., fluorescence, and

$$I(r, z) = \left| \left[ \text{PSF}_{\text{ill}}(r, z) \text{PSF}_{\text{det}}(r, z) \right] \otimes O(r, z) \right|^2, \quad (17.25)$$

for a coherent optical process. In the incoherent case which interests us here, the source is seen as a superposition of elementary point light sources. The coherent case is useful when treating other phenomena, such as nonlinear coherent emission mentioned in Sect. 17.3.2 (second harmonic generation SHG, coherent anti-Stokes Raman scattering CARS).

The underlying idea of confocal detection is to place the confocal diaphragm at the focal point where the image diffraction pattern is located, keeping only its central disk. The main advantage is the rejection of interfering light sources not belonging to the focal plane under investigation. As a consequence,  $\text{PSF}_{\text{det}}(r, z)$  is modified. The optical resolution thus obtained is not changed if the confocal diaphragm is much bigger than the Airy disk

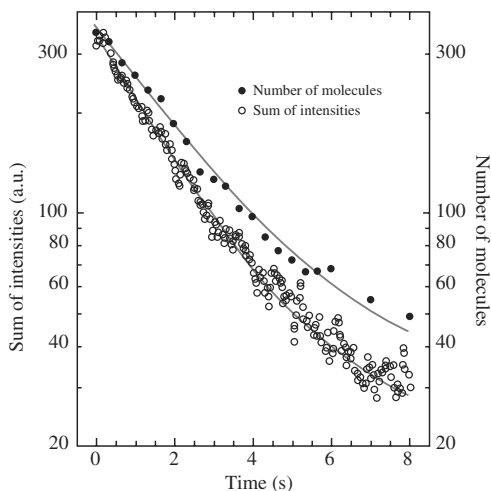


**Fig. 17.23.** Emission from single molecules immobilised within a polymer. Cy3 = isolated molecules of cyanine type. Bead = fluorescent reference sphere [35]

$\text{PSF}_{\text{ill}}(r, z)$ . When the diaphragm is much smaller than the Airy disk, diffraction effects due to the diaphragm must be taken into account in a form that can be solved analytically. The effect on the resolution is summed up in Fig. 17.22, where it can be seen that the longitudinal resolution is the one that is most affected by the diaphragm. We may conclude that the size of the confocal pinhole will be chosen in consequence of the compromise adopted between better luminosity and better longitudinal resolution. The size of the diaphragm is often chosen to be roughly equal to half the size of the Airy disk due to the objective  $\text{PSF}_{\text{ill}}(r, z)$ .

Figure 17.23 illustrates the characteristic emission of immobilised single molecules, together with a fluorescence image measured using confocal microscopy. Photobleaching is clearly visible in the emission time series for each of the molecules, as is the intermittent emission effect (blinking). From the series for a large number of molecules, the time development of an ensemble of molecules can be reconstituted. One retrieves the characteristic exponential decay of ensemble photobleaching (see Fig. 17.24). Note that similar observations have been made with regard to fluorescence emission by isolated CdSe/ZnS nanoparticles in a polymer matrix, mentioned in Sect. 17.1.

*Scanning Near-Field Optical Microscopy (SNOM).* Historically, near-field microscopy was the first technique used to detect isolated fluorescent molecules on a surface. This approach gradually gave way to confocal microscopy, which was easier to implement. Near-field microscopy involves illuminating the sample through a tapered optical fibre, producing a much higher resolution than



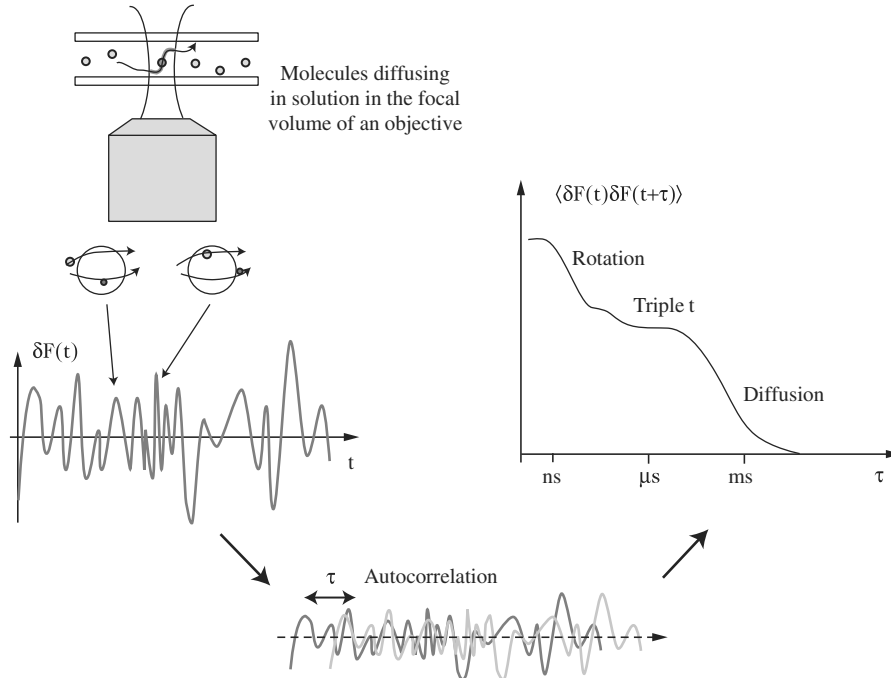
**Fig. 17.24.** Time emission from an ensemble of immobilised Cy3 molecules. Photobleaching of single molecules and ensemble measurements [35]

can be achieved by traditional fluorescence microscopy. Indeed, the diameter of the aperture at the end of the fibre can be as low as a few nanometers. In SNOM, optical measurements are based on a feedback loop involving the force felt by the probe (fibre). This system is similar to atomic force microscopy (AFM), except that the vibration of the optical fibre used as a tip reduces the resolution of the resulting images. This technique is discussed more fully in Chap. 5. When observing biological samples, the main difficulty today is to carry out high resolution measurements in aqueous media.

*Fluorescence Correlation Spectroscopy (FCS).* This technique consists in observing molecules that are freely diffusing in the focal volume of the microscope objective, but keeping the confocal configuration. The idea is to ensure that just one molecule is passing through this volume: the focal volume of an objective with large numerical aperture being  $10^{-15}$  L, the molecular concentration is reduced to  $10^{-9}$ – $10^{-12}$  mole/L. Very brief signals are then observed, typically in the range 0.1–1 ms. Several pieces of information can be deduced from such measurements: the diffusion time of the molecules (translational or rotational, if the detection is polarised), or the degree of excitation transfer between several fluorescence emitters (see Sect. 17.3.1), which is discussed below. Considering a time series  $F(t)$  for fluorescence emission, with time average  $\langle F(t) \rangle$  over a given integration time and fluctuations  $\delta F(t)$  about this average, the autocorrelation function  $G$  is given by

$$G(\tau) = \frac{\langle \delta F(t) \delta F(t + \tau) \rangle}{\langle F(t) \rangle^2}. \quad (17.26)$$

This function is generated by comparing the fluorescence emission in the focal volume at time  $t$  with the emission at time  $t + \tau$ , where  $\tau$  is the new time variable. This function contains a wealth of information concerning the sources of

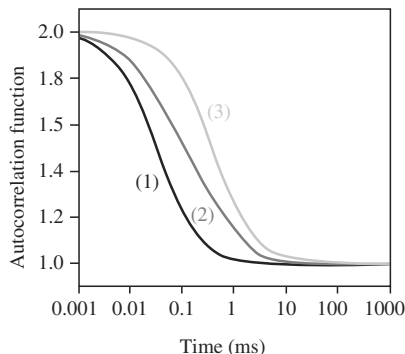


**Fig. 17.25.** Freely diffusing single molecules and schematic representation of the autocorrelation function

fluctuations in the emission from a single emitter, e.g., resulting from the interaction with its immediate environment, insofar as it succeeds in quantifying the various time scales of these fluctuations.

The measurement principle shown in Fig. 17.25 underlies the technique of fluorescence correlation spectroscopy (FCS), first introduced in the 1970s [36, 37]. This technique does not strictly speaking measure time series from a single molecule, but from the time series of several transiting molecules, it can provide information concerning their diffusion behaviour, the kinetics of chemical reactions, the molecular concentration, and other details.

It can be shown that, if the molecules diffuse freely in the focal volume with diffusion coefficient  $D_{\text{lat}}$ , then the autocorrelation function falls off exponentially with characteristic time  $\tau_{\text{lat}} = r_0^2/4D_{\text{lat}}$ , where  $r_0$  is the diameter of the Gaussian focusing beam measured at  $I_{\text{max}}/e^2$ . In a real situation, phenomena other than diffusion can be observed. At time scales shorter than the microsecond, behaviour such as (non-radiative) transitions into the triplet state or other photochemical processes become apparent, whereas longer time scales correspond to diffusion of the molecules through the focal volume. Observation on time scales shorter than the microsecond would reveal the rotational diffusion behaviour of the molecule. The translational diffusion coefficient  $D_{\text{lat}}$  is proportional to  $1/M^{1/3}$ , where  $M$  is the mass of the molecule, whereas the



**Fig. 17.26.** FCS of macromolecules. A fluorescent ligand is able to diffuse freely, giving rise to rapid fluorescence fluctuations, whereas a macromolecule to which the ligand had attached itself would diffuse less quickly, giving rise to slower fluctuations, and hence a higher level of correlation. As a consequence, the autocorrelation falls off more quickly in the case of free ligands (curve 1) than in the case of ligands bound to macromolecules (curve 3). A 1:1 mixture of free and bound ligands gives an intermediate curve (curve 2). Taken from [www.probes.com](http://www.probes.com)

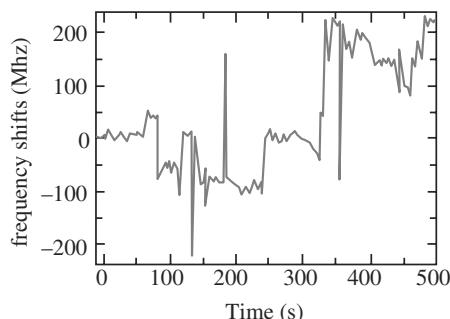
rotational diffusion coefficient  $D_{\text{rot}}$  is proportional to  $1/M$  [37]. The autocorrelation curves assume more complex shapes when chemical reactions with measurable kinetic coefficients are involved. Figure 17.26 shows how FCS can be used to distinguish biological entities.

*Low Temperature Spectroscopy of Single Molecules.* Until now, we have described how to detect single molecules dispersed in space. It is also possible to select a molecule among others by using the spectral characteristics of its absorption or emission, provided that we can distinguish it from the others. This is only possible at low temperatures, where the absorption and emission lines are narrow, or by using emitters with special properties, such as semiconductor nanocrystals. At low temperatures, effects that broaden emission lines such as collisions between molecules become negligible. The absorption line of a molecule at low temperature is thus reduced to its homogeneous width (in contrast to its inhomogeneous width in an ensemble of molecules at room temperature), and this is centered on a frequency which depends on the microscopic neighbourhood of the molecule. One can thus distinguish them in an ensemble where they occur frequently simply by scanning the wavelength of the excitation source.

There are other advantages in working at cryogenic temperatures. Photo-bleaching is reduced, since oxygen or water can no longer diffuse easily in the medium, and the absorption cross-section of the molecule is increased [25, 26, 28]. Indeed, for a randomly oriented molecule, it can be shown that

$$\sigma = \frac{\lambda^2}{2\pi} \frac{\gamma_{\text{r}}}{\Gamma_{\text{tot}}},$$





**Fig. 17.27.** Spectral shifts in the resonance frequency of a single pentacene molecule in a *p*-terphenyl crystal at 1.5 K. The spectral sequences are measured at 2.5-s intervals (origin 0 MHz = 592.546 nm). Changes in frequency, or spectral diffusion, are due to the reorientation of phenyl groups in molecules near the emitter [29]

where  $\lambda$  is the incident wavelength,  $\gamma_r$  is the radiative emission rate, and  $\Gamma_{\text{tot}}$  is the total absorption width (corresponding roughly to the sum of homogeneous and inhomogeneous widths in frequency). For a fluorophore in solution at room temperature,  $\gamma_r \approx 30$  MHz, which corresponds to a radiative lifetime of a few nanoseconds, and  $\Gamma_{\text{tot}} \approx 3 \times 10^4$  GHz. This returns the order of magnitude given earlier, viz.,  $\sigma \approx 6 \text{ \AA}^2$ . At low temperatures, the line width is reduced, e.g.,  $\Gamma_{\text{tot}} \approx 30$  MHz, whence  $\sigma \approx 10^6 \text{ \AA}^2$ , which corresponds to the cross-section of 100 000 molecules at room temperature!

At very low temperatures in a solid medium ( $< 10$  K), the homogeneous width of the molecular emission line is much narrower than the inhomogeneous width of their distribution ( $\gamma_{\text{hom}} \approx 10$  MHz). The effect of vibrational contributions is also absent. It is then possible to produce a spectral image of the molecules which provides useful information about their conformation or their interaction with their surroundings. In a crystalline environment, one can also observe significant spectral modifications (see Fig. 17.27).

Other effects have been observed at low temperatures, such as the bunching and antibunching of photons emitted by single molecules, and the shift in their emission line due to an electric field (Stark effect).

### Time Dynamics of Emission and Spatial Diffusion of Single Molecules

Important information can be obtained from the behaviour of isolated emitters by studying their emission dynamics, but also the polarisation state of the emitted light, or possible spectral fluctuations. These emitters are commonly used today as probes for their immediate environment. It thus becomes possible to study the dynamics of conformational change in macromolecules, or the interaction between proteins.

- In biology, a distinction is made between naturally fluorescing proteins [38], such as green fluorescent protein (GFP) or cholesterol oxidase (COx), and other systems which have to be marked (or functionalised) by fluorophores or nanocrystals in order to be able to observe them (see earlier sections). Marked systems may be proteins, DNA or RNA fragments, in aqueous media (gels), in artificial membranes (lipid bilayers or vesicles), in cell membranes (cell surfaces), or directly in a fixed or living cell [31–33, 38].
- In physicochemistry, many physical studies have also been carried out on chromophores included within polymer matrices or gels, in order to obtain information on heterogeneity and local behaviour in such media. Moreover, specific macromolecules display interesting emission properties, such as the antenna effect due to the large energy transfer within the molecule (luminescent polymers, dendrimers).

The following examples serve to illustrate single-molecule studies and the information it can yield.

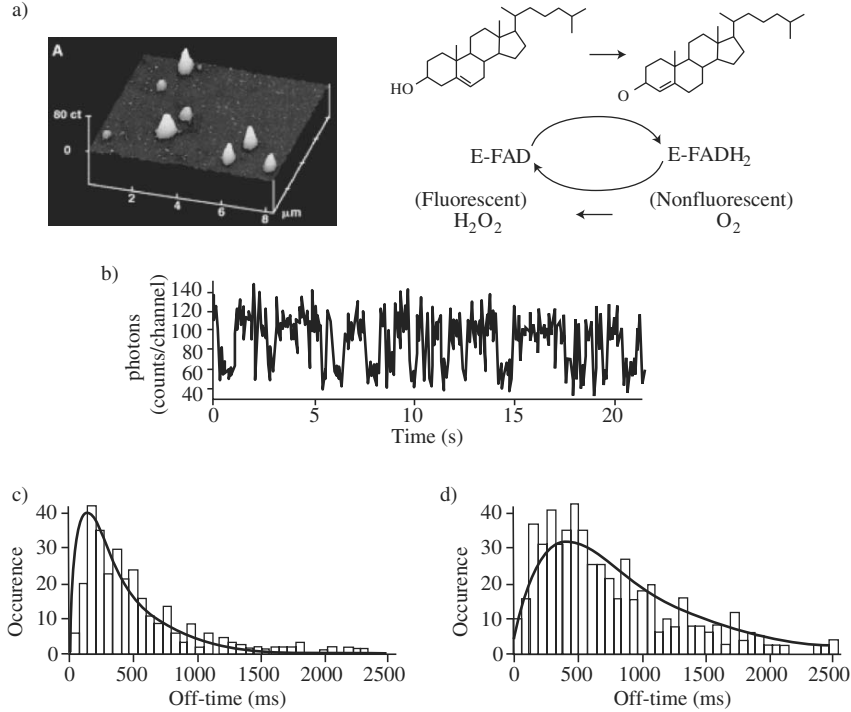
#### *Time Resolution: Emission Dynamics*

We saw earlier that the emission from a single molecule could be intermittent due to the various relaxation pathways towards non-radiative states. Intermittent emission may also arise through a change in conformation or chemical state of the molecule, with microsecond to second time scales to transit from one state to the other. It is thus possible to follow an equilibrium reaction between one emitting (on) state and one non-emitting (off) state of the molecule by measuring similar time series to those shown earlier. The mean time spent by the molecule in the on and off states provides direct access to the time scales of the relevant molecular dynamics. As an example, the behaviour of a specific protein is given in Fig. 17.28. This is cholesterol oxidase (COx), a flavoprotein enzyme which catalyses the oxidation of cholesterol by oxygen. The active site of the enzyme involves a flavin-adenine dinucleotide (FAD) which is naturally fluorescent in its oxidised form. It is reduced to a non-fluorescent form by cholesterol. The molecule thus passes successively from the fluorescent (on) state (oxidised) to the non-fluorescent (off) state (reduced) during the oxidation reaction.

The distribution of different behaviour measured over several molecules can thus be visualised, something that would be quite impossible through ensemble measurements.

#### *Polarisation Resolution: Reorientation Dynamics*

Earlier, we mentioned the dynamics of the emission from a single molecule. By monitoring the polarisation state of this emission, it is also possible to follow the real time evolution of the direction of the emitting dipole. The fluorescence signal of a single molecule is thus very sensitive to its orientation, represented by the direction of its emission dipole [33]. The fluorescence yield of a molecule



**Fig. 17.28.** (a) Image of immobilised cholesterol oxidase (COx) molecules, dispersed in an agarose gel and excited at 442 nm, with confocal detection at 520 nm. Oxidation of a COx molecule. (b) Real time observation of enzymatic reactions of a single COx molecule catalysing a cholesterol oxidation reaction (0.2 mM cholesterol in an agarose gel). The statistical distribution of on and off times depends on the amount of cholesterol: (c) 0.2 mM, (d) 2 mM [39]

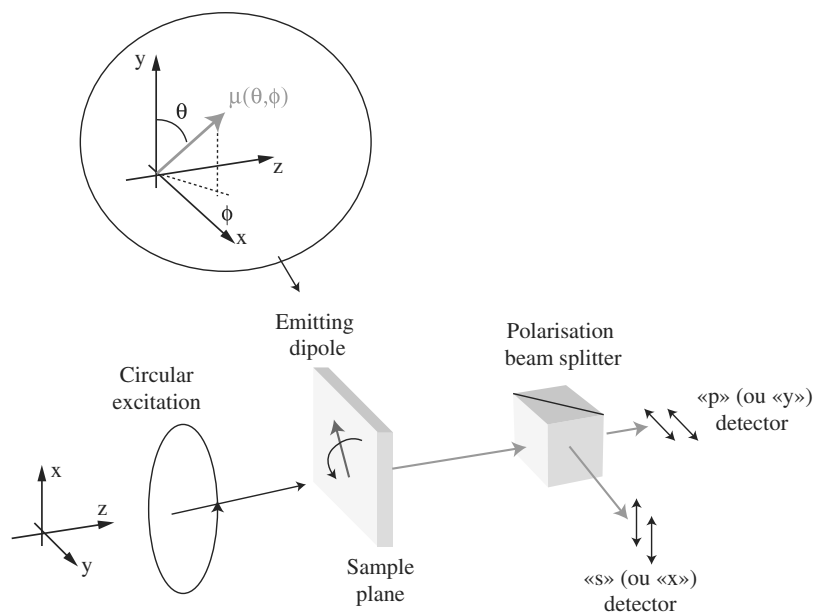
subjected to an optical field  $\mathbf{E}$  can be expressed in terms of the probabilities of absorption and emission, viz.,

$$P_{\text{abs}}(t) = |\boldsymbol{\mu}_{\text{abs}}(t) \cdot \mathbf{E}|^2, \quad P_{\text{em}}(t) = |\boldsymbol{\mu}_{\text{em}}(t) \cdot \mathbf{e}|^2, \quad (17.27)$$

as in (17.27) of Sect. 17.1, where  $\mathbf{e}$  is the polarisation direction, and  $\boldsymbol{\mu}_{\text{abs}}$  and  $\boldsymbol{\mu}_{\text{em}}$  are the absorption and emission dipoles of the molecule, respectively. The angle between the latter depends on the molecular conformation in its excited state (measurable by other means) [26]. Of course, the directions of these dipoles will evolve in time according to the dynamics of molecular rotation. The fluorescence signal thus has the form

$$F(t) \propto |\boldsymbol{\mu}(t) \cdot \mathbf{E}|^2 |\boldsymbol{\mu}(t) \cdot \mathbf{e}|^2,$$

for a molecule whose emission and absorption dipoles point in the same direction ( $\boldsymbol{\mu}_{\text{abs}} = \boldsymbol{\mu}_{\text{em}} = \boldsymbol{\mu}$ ), which is relatively common for long and dipolar molecules. By exciting a molecule with circular polarisation and simultaneously



**Fig. 17.29.** Setup for measuring dipole orientation by fluorescence

detecting the fluorescence emission along two perpendicular polarisations, its rotational diffusion dynamics can be monitored as a function of time. The point about using a circular incident polarisation is that one can excite the molecules with the same efficiency whatever their direction (see Fig. 17.29). The rotational diffusion time of molecules in a matrix depends heavily on the viscosity of their immediate environment and the possible interactions between molecules.

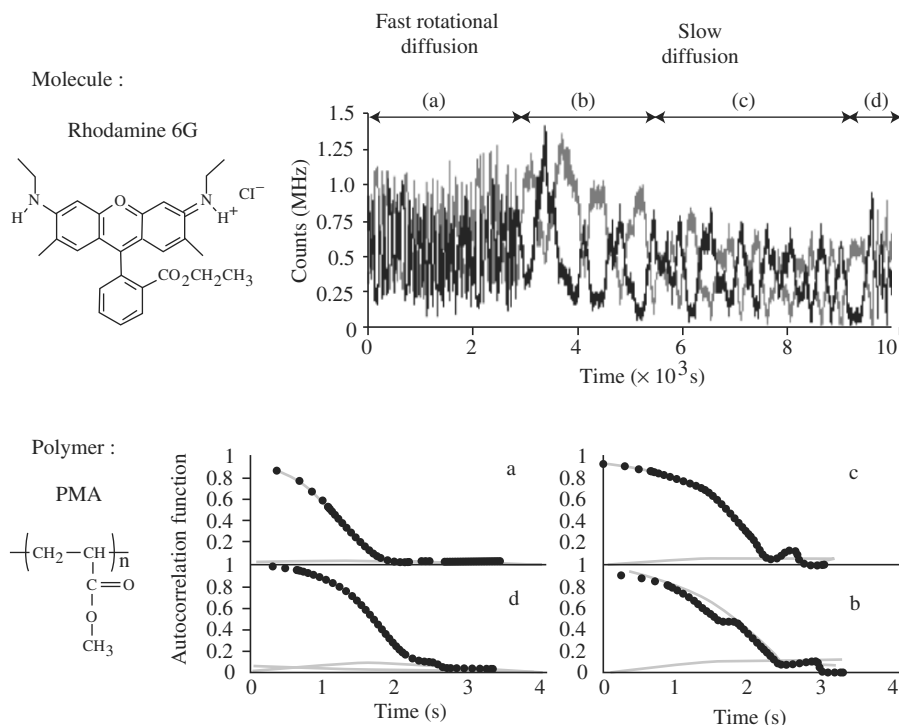
The orientation  $(\theta, \phi)$  of the dipole in space is defined by

$$\boldsymbol{\mu} = \mu(\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta) .$$

Using the notation of Fig. 17.29 and the above definition of the fluorescence signal, we then have  $F(t)_X \propto \sin^2 \theta(t) \cos^2 \phi(t)$  and  $F(t)_Y \propto \sin^2 \theta(t) \sin^2 \phi(t)$ . The anisotropy of the fluorescence signal depends directly on the angle of orientation in the plane and is defined by

$$r(t) = \frac{F(t)_Y - F(t)_X}{F(t)_Y + F(t)_X} = \cos 2\phi(t) . \quad (17.28)$$

By measuring the fluorescence anisotropy of a single molecule, one thus directly obtains its orientation projected in the sample plane, the value being averaged over the integration time of the detectors. In the example of Fig. 17.30, it can be seen that the orientations of molecules inserted by dilution into the pores of a polymer are able to evolve rapidly, and that changes of regime are possible as time goes by, depending on the very dynamics of the immediate



**Fig. 17.30.** Emission dynamics along two orthogonal polarisations (*grey* = *p*, *black* = *s*) of a single Rhodamine 6G molecule immobilised with respect to lateral diffusion in a poly(methyl)acrylate (PMA) polymer film, 10 K above its glass transition temperature (287 K) [27]

surroundings with which they interact. A change in the rotational dynamics is largely reflected by modifications in local properties like the pH and the viscosity.

The local viscosity causes constraints that prevent the molecule from diffusing freely in the matrix, constraints that could not be revealed by ensemble measurements. This example illustrates the usefulness of single-molecule spectroscopy when it comes to understanding heterogeneity in the behaviour of a matrix that is a priori homogeneous. The various rotational diffusion regimes shown in Fig. 17.30 can be interpreted quantitatively by examining the temporal autocorrelation of the emission. This function exhibits a multiexponential decay indicating the presence of several different rotational regimes. However, each regime taken separately (on shorter time scales, see Fig. 17.30) exhibits a monoexponential decay. This observation shows that a single molecule is sensitive to dynamical disorder in the matrix, in such a way that over a long period of time the autocorrelation measurement begins to look like what would be expected from the dynamics of an ensemble measurement. This agrees with the

ergodic principle concerning the behaviour of these isolated molecules: observing an average signal over a large number of molecules amounts to observing the behaviour of a single one of these molecules over a certain time.

*Excitation Transfer: Conformation Dynamics*

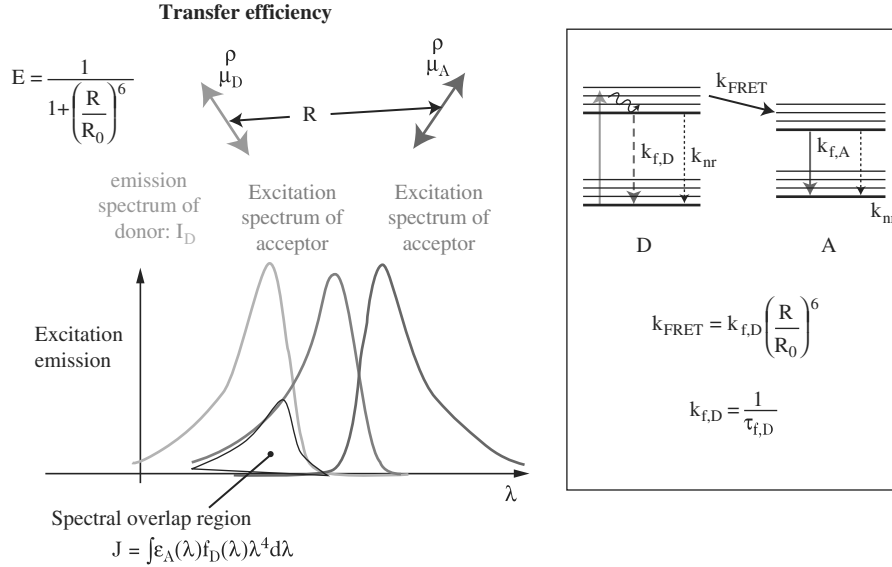
There is a special case of the interaction between a molecule and its environment that is of particular interest, in which the molecule find itself at a very short distance ( $< 10\text{--}100 \text{ \AA}$ ) from another emitting dipole. In this situation, an excitation transfer may occur. This is called fluorescence resonance energy transfer (FRET), originally identified by Förster. For an emitter (donor D) in close proximity to another (acceptor A), and if the absorption spectrum of A overlaps with the emission spectrum of D, part of the energy absorbed by D is transferred non-radiatively by dipole–dipole interaction to A with an efficiency  $E$  given by

$$E = \frac{1}{1 + (R/R_0)^6}, \quad (17.29)$$

where  $R$  is the distance between A and D, and  $R_0$  is the Förster radius. The latter is such that  $R_0^6 \propto \Phi_D J \kappa^2 / n^4$ , where  $\kappa$  is a geometric term such that  $\kappa = \cos \theta_{AD} - 3 \cos \theta_{AR} \cos \theta_{DR}$ , taking into account the respective orientation of the two dipoles and the vector  $\mathbf{R}$  joining them (defined by their direction cosines  $\cos \theta_{AD}$  and  $\cos \theta_{AR} \cos \theta_{DR}$ ),  $n$  is the refractive index of the medium, and  $\Phi_D$  is the fluorescence quantum yield of the donor D. The quantity  $J$  is the spectral overlap integral for the absorption of D and the emission of A (see Fig. 17.31). Note that, when the molecules diffuse freely with a small rotational diffusion time compared with the integration time of the detectors, the time average of  $\kappa^2$  is equal to  $2/3$ . In this situation, one does not need to measure the orientations of the dipoles over time.

The length  $R_0$ , typically a few nanometers, corresponds to the distance at which A must be located to obtain a transfer efficiency of 50%. Excitation transfer between two fluorescent dipoles is therefore a phenomenon capable of providing information on the distance between two distant emitters with subnanometer resolution.

This effect, commonly used for ensemble measurements in solution or biological medium, can be applied in the single-molecule context for its sensitivity to the distance between two emitters on macromolecular scales. For example, it can inform about conformational changes of an isolated protein by functionalising at two different sites of this protein, using one molecule to play the part of donor and another to play the part of acceptor. Applications encompass a variety of effects, including fluctuation and stability of macromolecular conformation, folding/unfolding dynamics of a protein, and structural changes in an enzyme during catalysis (see Fig. 17.32). Research into protein folding is particularly appropriate on the level of single molecules, given its complexity and stochastic nature which make it difficult to interpret ensemble measurements. In such conditions, various paths of conformational change can in fact



**Fig. 17.31.** Excitation transfer between donor and acceptor dipoles. The transfer is defined by its rate  $k_{\text{FRET}}$  ( $\text{s}^{-1}$ ), which provides a further relaxation channel for the donor D, competing with its fluorescence emission rate  $k_{f,D}$  ( $\tau_{f,D}$  is the fluorescence lifetime of D)

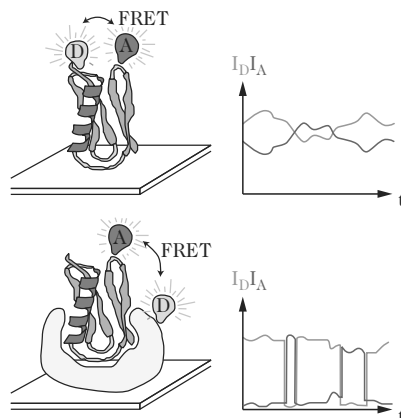
be identified, along with their characteristic time scales. This information is crucial given the relevance of protein conformation to the way cells work.

In practice, the signals corresponding to the emission wavelengths of the donor and acceptor can be measured in confocal microscopy for a given excitation wavelength, by using suitable spectral filters. The signals are then corrected for the quantum yields and collection efficiencies for the donor and acceptor. One then measures

$$E = \left(1 + \alpha \frac{F_D Q_A}{F_A Q_D}\right)^{-1},$$

where  $F_A$  and  $F_D$  are the fluorescence signals of A and D, respectively,  $Q_A$  and  $Q_D$  are the fluorescence quantum yields, and  $\alpha$  is an experimental correction coefficient accounting for the collecting efficiencies of the two detection channels. Figure 17.32 is a schematic representation of two different situations in which donor and acceptor molecules are attached to a large molecule, together with the expected signals.

We can illustrate the measurement of a FRET signal in single-molecule spectroscopy by the example of the protein chymotrypsin inhibitor 2, a simple model system that has two equilibrium states: a folded conformation and an unfolded conformation. The two stable conformation states of the protein have been revealed by exploring intermediate situations in which increasing



**Fig. 17.32.** Illustration of energy transfer in single-molecule detection. *Upper:* Folding dynamics of a protein revealed by following the signals from two markers attached to the protein. *Lower:* Interaction of an enzyme with the substrate.  $I_D$  photodiode detecting the fluorescence signal at the donor wavelength.  $I_A$  photodiode detecting the fluorescence signal at the donor wavelength [7]

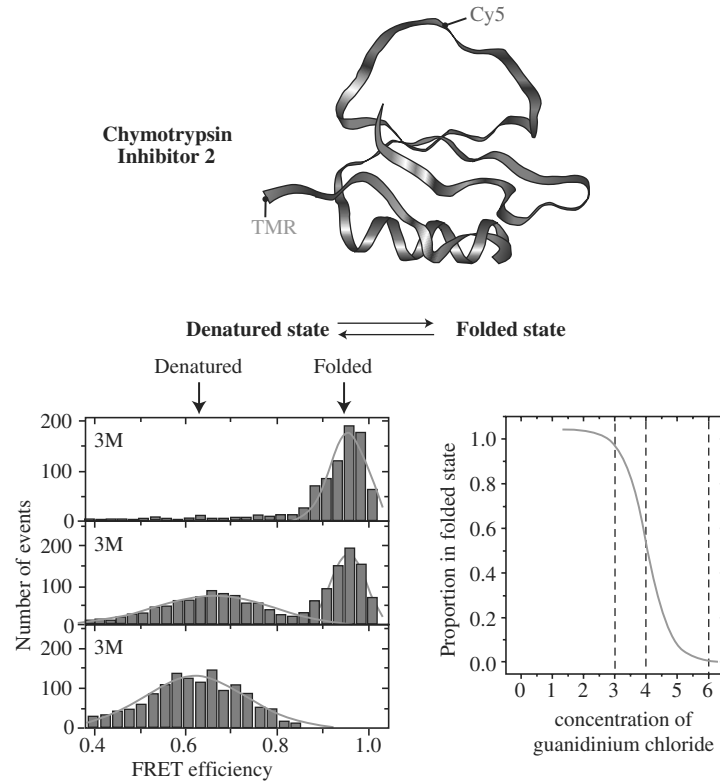
concentrations of a denaturing agent (guanidinium chloride) have been added. The most compact state, bringing the two molecules close to one another (45 Å, ascertained from the transfer efficiency), produces a strong FRET signal. The other state increases the separation between the two molecules to 61 Å. The measurements shown in Fig. 17.33 were made on molecules diffusing freely in the focal volume of a confocal microscope. This measurement can provide proof that these two states do exist. Other experimental setups are possible, involving immobilisation of the molecules, for example, which allows one to study the dynamics of transitions between the various conformations.

#### *Spatial Resolution: Diffusion in Membranes*

The possibility of imaging fluorescence areas measuring a few tens of microns with high optical resolution ( $\sim 300$  nm) and a high level of sensitivity has opened the way to detecting single molecules that are diffusing, freely or otherwise, in fluid environments such as polymers and artificial membranes like lipid bilayers, cell membranes, and cells [43]. The diffusion behaviour of single molecules in a given environment provides a wealth of information about its interaction with its close surroundings, e.g., protein–lipid, protein–protein, protein–cytoskeleton in a cell membrane. Specific behaviour such as constrained diffusion, which would be hidden in an ensemble measurement, may also be revealed.

Figure 17.34 shows the cell components which may be relevant in protein, DNA, or virus diffusion properties at the level of the single molecule. Membrane proteins diffuse at the surface in a fluid medium (membrane, lipid

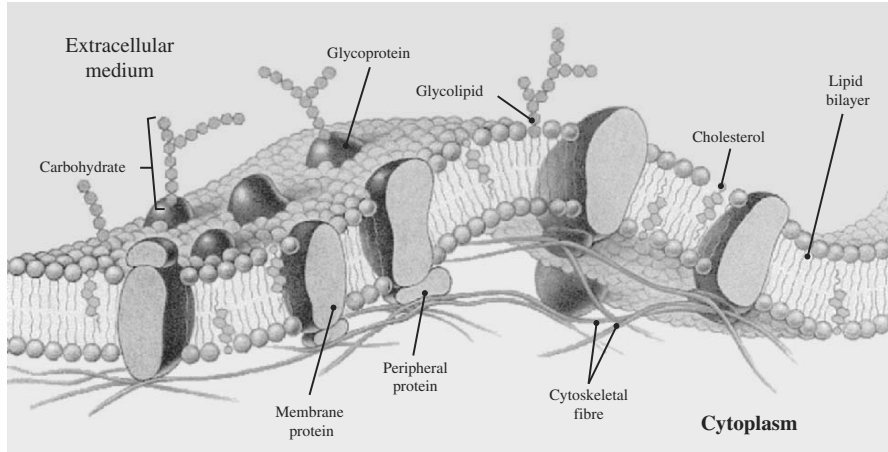




**Fig. 17.33.** Stable conformation states of the protein chymotrypsin inhibitor 2 investigated by energy transfer between the dye TMR ( $\lambda_{\text{emission}} \approx 580 \text{ nm}$ ) and the dye Cy5 ( $\lambda_{\text{excitation}} \approx 640 \text{ nm}$ ), attached to two remote sites on the protein. The equilibrium between the folding states of the protein is shifted by increasing the concentration of a denaturing agent from 3 mole/L to 69 mole/L [41]

bilayer) comprising many obstacles and specific domains which may play a specific role in cell recognition, for example. Diffusion through the cell membrane, an essential phenomenon with regard to cell nutrition and waste rejection by the cell, occurs either by free transport, i.e., diffusion from zones with the highest molecule concentrations to less concentrated regions, or by active transport, i.e., the cell 'holds' the molecules in spite of gradient concentrations, thereby maintaining an equilibrium with its surroundings. Active transport is achieved by specific membrane proteins which use cell energy supplied by adenosine triphosphate (ATP). This is a molecule which plays a determining role with regard to energy supply by transferring its terminal phosphate group directly to the protein that acts as intermediary for the transport.

The theoretical approach to particle diffusion is based on a hydrodynamic model. Historically, the first demonstrations were obtained by observing the diffusion of pollen particles in water (R. Brown in 1827, then A. Einstein in



**Fig. 17.34.** Cell membrane. Boundary between the inside of the cell and its various components and the outside. Taken from [42]

1905 and P. Langevin in 1908). The general equation of motion

$$m \frac{dv}{dt} = -\alpha v + F(t)$$

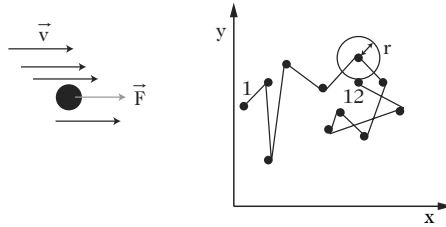
gives the velocity of a particle subjected to a fluctuating force  $F(t)$ , such that  $\langle F(t) \rangle = 0$ , and friction with coefficient  $\alpha$  expressing dissipation. To study diffusion of molecules, macromolecules and particles in cell membranes or within a cell, we limit the discussion to 2D diffusion in a plane, although the properties described here can be extended to three dimensions [44]. For purely Brownian diffusion in two dimensions, the spatiotemporal variation of the probability of finding a particle in a given environment is given by Fick's law:

$$\frac{\partial P(\mathbf{r}, t)}{\partial t} = D \nabla_{\mathbf{r}}^2 P(\mathbf{r}, t), \quad (17.30)$$

where  $P(\mathbf{r}, t)$  is the probability of the particle moving to the position  $\mathbf{r}$  at time  $t$ , measured from its original position. The amplitude of the displacement  $\|\mathbf{r}\|$  is indicated in Fig. 17.35, which shows an example of position measurements on a particle diffusing in the plane.  $D$  is the lateral diffusion coefficient, given by  $D = 3kT/\alpha$  according to the Einstein model mentioned above. The solution to this equation is a probability density of the form

$$P(\mathbf{r}, t) d\mathbf{r} = \frac{1}{\sqrt{8\pi Dt}} \exp\left(-\frac{r^2}{4Dt}\right) d\mathbf{r}.$$

The function most often used directly relates the probability of finding the particle at a given distance from its previous position:



**Fig. 17.35.** Diffusion of particles in a plane. The position of the particle is measured at regular time intervals  $\Delta t$

$$P(r^2, t) = 1 - \exp\left(-\frac{r^2}{4Dt}\right). \quad (17.31)$$

This function can be plotted directly from measurements like those represented in the figure. If the theoretical parameters can be fitted to the experimental data to yield the exponential decline given in this equation, the particle motion can be considered to be Brownian.

In practice, it is also possible to make direct measurements of molecular positions every  $n\Delta t$  and record the mean squared displacement defined by

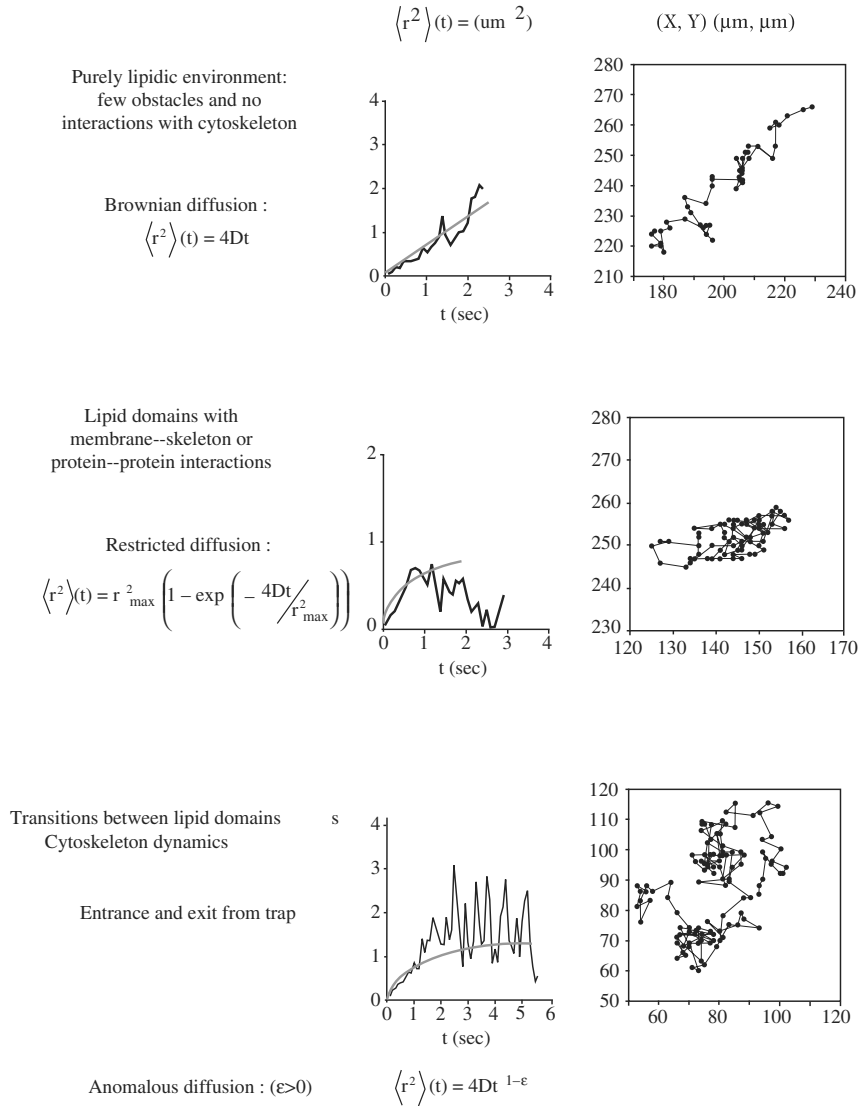
$$\langle r^2 \rangle (t = n\Delta t) = \frac{\sum_{i=0}^N [(x_{i+n} - x_i)^2 + (y_{i+n} - y_i)^2]}{(N + 1)}, \quad (17.32)$$

where  $x_i$  and  $y_i$  are the particle position coordinates as measured at time  $i\Delta t$ .

This approach gives experimental access to the time-dependent particle displacement. This function is known in the simple cases described in Fig. 17.36. Most of these situations (diffusion restricted to domains of finite dimension, anomalous diffusion) have been the subject of much theoretical and experimental work, leading to empirical models that are widely used today. For purely Brownian motion, as observed in the diffusion behaviour of proteins in an artificial lipid membrane, the mean squared displacement is a linear function of time, viz.,  $\langle r^2 \rangle = 4Dt$ , and the coefficient  $D$  can be deduced directly. Any deviation from Brownian behaviour is thus directly identifiable.

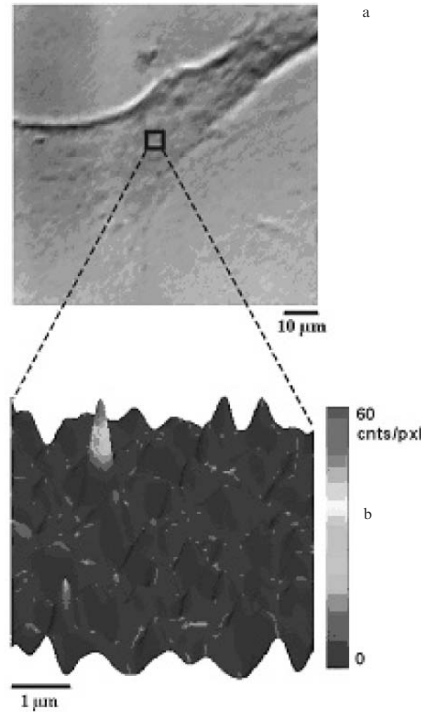
More complex models have been developed to explain the diffusion behaviour of membrane proteins, taking into account the local viscosity of the lipid environment, for example (Saffman-Delbrück, 1975).

In practice, measurements of the position of a molecule depend on the optical resolution, but the displacement can be estimated with a resolution of around 50 nm. Typically, a membrane protein diffusing freely in a cell membrane diffuses with diffusion constant  $D \approx 5\text{--}10 \mu\text{m}^2/\text{s}$ . Under the experimental conditions of confocal microscopy, a particle is considered to be motionless for  $D < 10^{-4} \mu\text{m}^2/\text{s}$ . The signal-to-noise ratio is a very important factor for the detection of single molecules in the cell medium. Indeed, the cell environment is itself made up of fluorescing entities such as flavins, proteins that will fluoresce under excitation at 500 nm. To remedy this difficulty, excitation in the red is favoured, e.g., HeNe laser at 632.8 nm (see Fig. 17.37), or two-photon excitation (see below).



**Fig. 17.36.** Diffusion behaviour in a complex medium. These examples are accessible to fluorescence observations, by measuring the displacement of isolated proteins labelled by a fluorophore. Taken from experimental data on Cy5 molecules marking MHC membrane proteins on Chinese hamster ovary (CHO) cells [45]

More detailed studies of protein diffusion within a cell have shown that it is now possible to detect intracellular markers, as in the example of Fig. 17.38, where a labelled virus has been followed from the moment it crossed the membrane (endocytosis) until it reached the nucleus. In this example, it was possible to determine the very nature of the infection process by monitoring



**Fig. 17.37.** Detection of single molecules in a cell membrane. (a) Transmission electron microscope image of an HASM cell (adhering to the substrate) in which lipid probes (DOPE, unsaturated lipid 1,2-dioleoyl-sn-glycero-3-phosphoethanolamine) have been marked in very small quantities by a fluorescent cyanine (Cy5). The cell is imaged with magnification  $\times 40$  in (a), whereas the fluorescence image (b) was recorded with a magnification of  $\times 100$ . A Cy5 is clearly visible in the latter, with a signal-to-noise ratio of 23. The diffusion of DOPE-Cy5 is Brownian with coefficient  $D \approx 3.0 \mu\text{m}^2/\text{s}$ , whereas observations of the diffusion of marked saturated lipids DMPE-Cy5 [(1,2-dimyristoyl-sn-glycero-3-phosphoethanolamine)-Cy5] exhibit constrained diffusion behaviour in domains of measurable size: diffusion constant  $D \approx 0.6 \mu\text{m}^2/\text{s}$  and domain sizes  $700 \pm 20 \text{ nm}$  [46]

the intracellular diffusion of the virus. In particular, understanding the transport phenomena exploited by the virus to reach the nucleus (where the genes are expressed) would provide a way of identifying the decisive steps to target in the development of a suitable therapy.

### 17.3.2 Multiphoton and Nonlinear Microscopy

The aim in this section is to describe the main techniques of multiphoton fluorescence microscopy (two- or three-photon excitation), and nonlinear coherent microscopy (second harmonic generation SHG, third harmonic generation



**Fig. 17.38.** Fluorescence imaging of the adeno-associated virus (AAV) marked by a fluorophore (Cy5) and diffusing within a HeLa cell. The virus concentration has been reduced in order to isolate them. The transmission electron microscope image has been superposed on the image of the trajectory of various isolated viruses as monitored by fluorescence. Trajectories 1 to 4 show different stages of the infection. (1 and 2) Diffusion in solution outside the cell. (3) Penetration into the cell membrane. (3 and 4) Diffusion in the cytoplasm. (4) Penetration in the nuclear envelope and diffusion in the nucleoplasm. Diffusion is free (passive transport) and sometimes anomalous in cases 3 and 4 [47]

THG, coherent anti-Stokes Raman scattering CARS). Most of these techniques are still at the research stage.

### Two-Photon Fluorescence

Multiphoton fluorescence microscopy has been presented recently as a useful alternative to confocal microscopy. The excitation wavelengths are longer, in the near infrared, thereby reducing light scattering effects in the kind of complex media represented by cells and increasing the depth of penetration in consequence. Photodamage is also reduced in the infrared, and so is the fluorescence background due to the cell environment itself.

In two-photon fluorescence microscopy, excitation occurs spectrally in the infrared, typically in the range 750–1000 nm, since two photons must be absorbed to generate fluorescence in the visible. Because the excitation requires two photons, incident energies are also greater than those used in conventional or confocal one-photon fluorescence microscopy. To this end, pulsed lasers are used, with subpicosecond pulse width (generally 30–200 fs) and high repetition rates (typically 80–100 MHz), leading to average power outputs in the range 30–300 mW. Optical configurations for two- or three-photon fluorescence remain essentially the same as in confocal microscopy (see Sect. 17.3.1). However, the confocal pinhole is no longer necessary, because the observed

effects are sensitive to the second or third power of the incident intensity, so that the observed phenomenon has an intrinsic longitudinal resolution (see Fig. 17.39).

The main difference with one-photon fluorescence is the excitation process. The excitation probability in two-photon fluorescence is proportional to the fourth power of the field, whence the measured intensity is proportional to [see (17.8)]

$$I^{2\text{-photon}} = |\boldsymbol{\mu}_{\text{exc}} \cdot \mathbf{E}|^4 |\boldsymbol{\mu}_{\text{em}} \cdot \mathbf{e}|^2 .$$

One may thus define a two-photon absorption cross-section, written  $\delta$  (in units of  $\text{cm}^4\text{s}/\text{photon}$ ).

#### *Resolution with Two-Photon Fluorescence*

The reduction in the interaction volume is due to the reduction in the intensity distribution PSF defined above. Hence, for a fluorescence process of order  $N$ , the new PSF, denoted by  $\text{PSF}^{(N)}$ , will be related to its equivalent linear PSF by  $\text{PSF}^{(N)} = (\text{PSF})^N$ . It can be shown that, for a Gaussian distribution, this induces a division of the original size by  $\sqrt{N}$ . For two-photon fluorescence, one may write

$$I^{2\text{-photon}} = |\text{PSF}_{\text{ill}}|^4 |\text{PSF}_{\text{det}}|^2 ,$$

whereas

$$I^{1\text{-photon}} = |\text{PSF}_{\text{ill}}|^2 |\text{PSF}_{\text{det}}|^2 .$$

In contrast to one-photon fluorescence as described in Sect. 17.1, the signal emitted in two-photon fluorescence is now expressed as a function of the square of the incident intensity integrated over the focal volume  $V$  [49]:

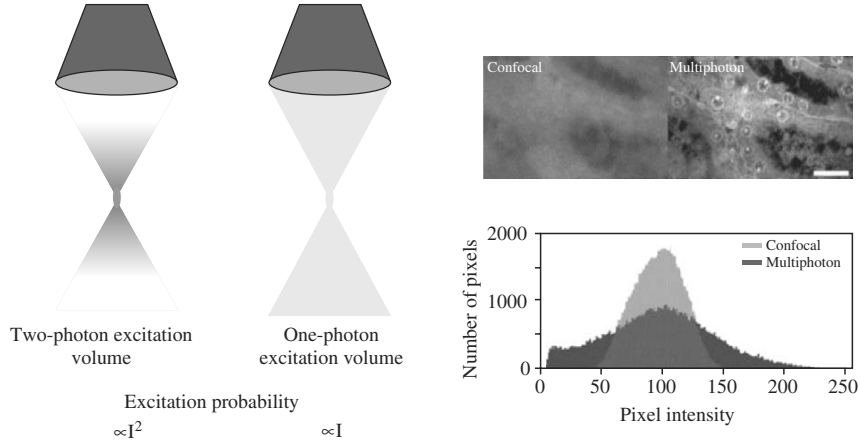
$$\langle F(t) \rangle^{2\text{-photon}} = CQ^{(2)} \langle I_0^2(t) \rangle \delta \int_V a^2(r) d^3r , \quad (17.33)$$

where  $a(x)$  is the spatial distribution of excitation (dimensionless),  $Q^{(2)}$  is the two-photon fluorescence quantum yield,  $I_0(t)$  is the incident intensity,  $\delta$  is the two-photon absorption cross-section, and  $C$  is an experimental coefficient representing the collection efficiency.

The expression for  $\int_V a^2(r) d^3r$  can be used to deduce the spatial resolution of the process. After spatial integration of a homogeneous profile (see above in the case of conventional microscopy), we have

$$\int_V a^2(r) d^3r \propto \frac{8n\lambda^3}{\pi^3 \text{NA}^4} ,$$

where NA is the numerical aperture of the objective and  $n$  is the refractive index of the medium. The diffraction-limited focal volume is thus proportional



**Fig. 17.39.** *Left:* Excitation volume for one- and two-photon fluorescence. *Right:* Confocal and two-photon fluorescence images of a tissue (at a depth of  $60\ \mu\text{m}$  in a monkey kidney labelled by a fluorophore) requiring a large penetration depth [48]

to  $\lambda^3/\text{NA}^4$ , in contrast with the case of one-photon fluorescence for which the dependence is  $\lambda^3/\text{NA}^3$ . Volumes of  $0.06\ \mu\text{m}^3$  are typical under two-photon excitation.

The spatial integration leads to the following simple expression, which takes into account the quantity  $\langle I_0(t) \rangle^2$ , directly measurable by a detector:

$$\langle F(t) \rangle^{2\text{-photon}} = CQ^{(2)} \langle I_0(t) \rangle^2 \delta g \frac{8n\lambda^3}{\pi^3 \text{NA}^4}, \quad (17.34)$$

where  $g = \langle I_0^2(t) \rangle / \langle I_0(t) \rangle^2$  is a measure of the temporal coherence of the source ( $g = 1$  for a CW laser source). For a pulsed laser with repetition rate 100 MHz and Gaussian pulse profile of width 100 fs, one has  $g \sim 10^5$ .

*Two-Photon Fluorescence Imaging*

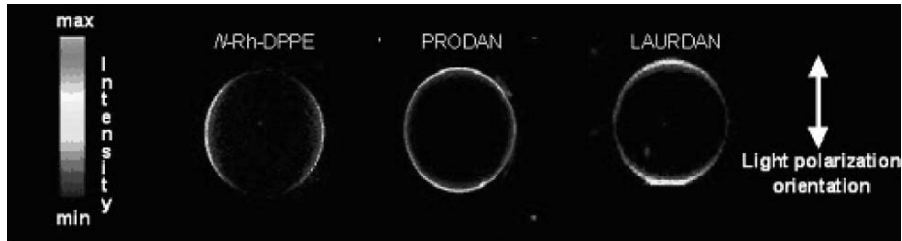
Two-photon fluorescence is particularly well-suited to 3D imaging inside a cell, where access to inner regions is more easily achieved. This feature is illustrated in Fig. 17.40, where labelled slices of vesicles have been imaged by polarised two-photon fluorescence microscopy.

**Two- and Three-Photon Coherent Phenomena:  
Second and Third Harmonic Generation**

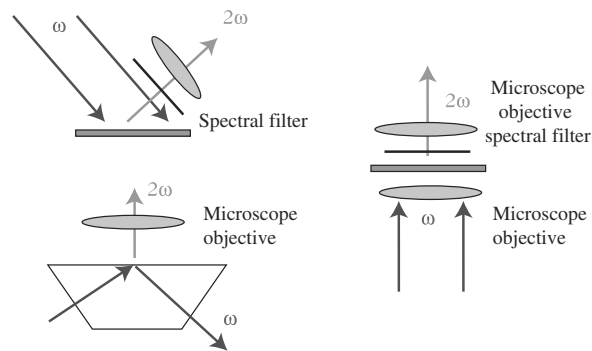
*Second Harmonic Generation (SHG)*

The phenomenon of second harmonic generation was described in Sect. 17.1. In microscopy, it is possible to detect the signal generated in molecular ensembles with non-centrosymmetric distribution either by transmission (a second





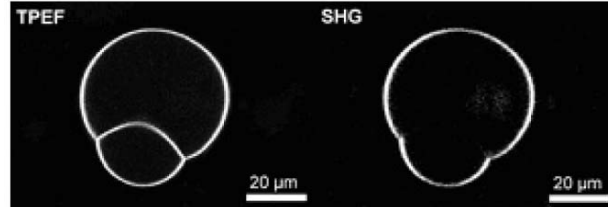
**Fig. 17.40.** Two-photon fluorescence imaging. Giant unilamellar vesicles (GUV) are model systems of cell dimensions which are suitable for studying lipid–lipid and lipid–protein interactions. These vesicles occur in the form of lipid membrane spheres and are easily labelled by fluorescent dyes, whose orientation depends on the dye structure. Two-photon fluorescence can be used to produce images of vesicular slices. The polarisation response of the generated signals provides information concerning the structure of the vesicle, its phase, etc. [50]



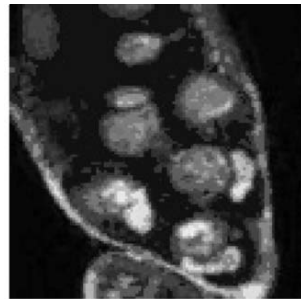
**Fig. 17.41.** Different setups for detection in SHG microscopy

objective is then necessary), or by reflection (see Fig. 17.41). This technique is particularly well-suited to the study of morphological changes in a biological medium [53], the detection of defects in materials, and the analysis of specific surface properties. The interpretation of images in terms of molecular organisation can be achieved by polarisation analysis. Recently, SHG microscopy has also been extended to the near-field configuration in molecular media [53].

The following example, shown in Fig. 17.42, illustrates the complementarity of information provided by two-photon fluorescence and coherent SHG emission on the same system, using the same optical excitation in the infrared. Giant vesicles (GUV) have been doped with nonlinear fluorescent molecules, and it can be seen that the fluorescence is visible over the whole surface, whereas the SHG disappears in the interface region between two vesicles. This is due to the fact that, unlike the SHG, fluorescence exists even in a centrosymmetric medium. In the interface region where the molecules lie head-to-foot, the order loses its centrosymmetry and this cancels the SHG signals.



**Fig. 17.42.** Two-photon fluorescence (*left*) and SHG (*right*) microscopy in vesicles doped by lipid dyes with non-negligible nonlinear quadratic and two-photon fluorescence yields. The part of the image where the two vesicles fuse is inactive in SHG because, in this region of the lipid layer, the molecules are oriented head-to-foot [54]



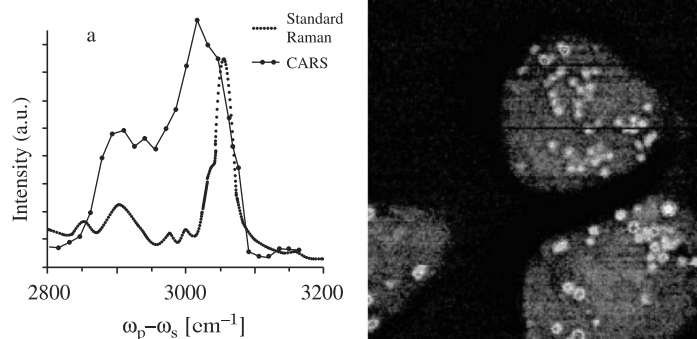
**Fig. 17.43.** THG (and SHG) microscopy in a living cell from a drosophila ovary labelled by a dye (DAPI) [55]

### *Third Harmonic Generation*

Whereas SHG requires a medium with a centre of symmetry, i.e., that is not centrosymmetric, the process of third harmonic generation (THG) is allowed in any medium. (It is in fact a process of odd order, as described in Sect. 17.1.) The THG signal, emitted at frequency  $3\omega$ , where  $\omega$  is the incident frequency, is proportional to  $|\chi^{(3)}|^2 I(\omega)I(\omega)I(\omega)$ , and hence proportional to the third power of the incident intensity. This optical process, which requires high incident energies, has nevertheless proven its efficiency for observing changes of phase in liquid crystals, optical fibres, and indeed biological media.

### **Coherent Anti-Stokes Raman Scattering (CARS)**

In a similar way to third harmonic generation, coherent Raman emission is a third order process (or a four-wave mixing process), which is in fact resonant for vibrational transitions rather than electronic levels. This process is therefore very useful for spectroscopic detection in which the molecules and molecular bonds can be directly identified spectrally. In the CARS process, two incident frequencies  $\omega_p$  (pump) and  $\omega_s$  (signal) interact, with  $\omega_p - \omega_s$  of



**Fig. 17.44.** CARS microscopy on polystyrene beads (*left*) and on a living cell (*right*) [57]

the order of a resonance frequency of the medium. The third order nonlinear polarisation induced at frequency  $\omega_{AS} = 2\omega_p - \omega_s$  is proportional to

$$\mathbf{P}_{\omega_{AS}}^{(3)} = \chi^{(3)} : \mathbf{E}(\omega_p)\mathbf{E}(\omega_p)\mathbf{E}(\omega_s)^* .$$

The latter is the source of radiation of the detected signal. The point about using CARS microscopy is that one can work on the molecular level in biological media such as cells without needing to inject fluorescent dyes [56]. The subjects of study are identified by their Raman resonance frequency. This nonlinear process can also be used to shift the detection frequency far from the molecular fluorescence line, which can otherwise lead to optical noise problems in classical Raman measurements. The first demonstrations on polystyrene beads confirmed the feasibility of such a technique (see Fig. 17.44).

### 17.3.3 Mechanical Properties of Single Biomolecules

#### Optical Tweezers

Optical tweezers are used to hold a microscale object under a focused light beam and to measure the effects of displacing this object under the action of extremely small forces (1–100 pN). By attaching them to a molecule or biomolecule, optical tweezers can be used to measure the forces applied to this molecule, or to exert forces with a view to deforming or displacing it.

Optical tweezers use forces resulting from radiated light pressure (defined below) to trap small particles. This technique has been used over the past 20 years to trap dielectric particles of micrometric dimensions, and it is only recently that force measurements have been carried out on molecules in biological environments, manipulating them with nanoscale spatial accuracy [32, 58–60]. This new type of manipulation of single molecules opens the way to a wide range of associated studies, e.g., to test the physical behaviour of DNA (conformational elasticity) from a standpoint that was never before

accessible. Likewise, many effects have now been understood in the field of biological molecular motors, and interaction phenomena between proteins and DNA by direct measurement of the interaction energy during binding. Several examples are outlined below.

### *Optical Traps*

When a light beam crosses a dielectric particle, the directions of the rays in the beam are modified by refraction. This causes a change in the direction of the momentum  $\mathbf{p}$  of the photons, and hence a radiation pressure on the particle, defined as the force per unit area exerted by the momentum  $\mathbf{p}$  of the photons. The photon momentum is given by  $\mathbf{p} = \hbar\mathbf{k}$ , where  $\mathbf{k}$  is the wave vector with amplitude  $|\mathbf{k}| = 2\pi/\lambda$  for a beam of wavelength  $\lambda$ . The requirement whereby the total momentum of the electromagnetic wave and the particle is conserved allows one to express the resulting force on the particle, this being given as the difference between the incoming light flux at the particle (of area  $\Sigma$ ) and the outgoing light flux:

$$\mathbf{F} = \frac{n}{c} \iint_{\Sigma} (\mathbf{S}_{\text{in}} - \mathbf{S}_{\text{out}}) dA, \quad (17.35)$$

where  $\mathbf{S}$  is the Poynting vector related to the total momentum of the photons by

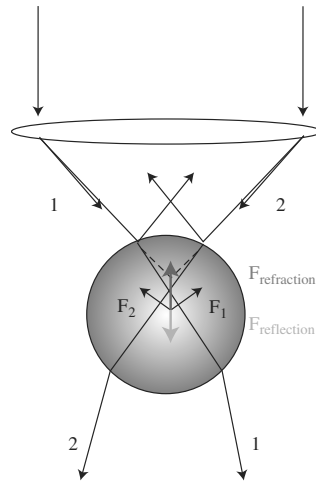
$$\frac{d^2\mathbf{P}}{dt^2} = \frac{n}{c} \mathbf{S} dA.$$

This force is generally very weak compared with other forces, such as the weight. This is why the objects manipulated must be very small, e.g., polystyrene spheres with diameters of the order of a few nanometers to a few microns.

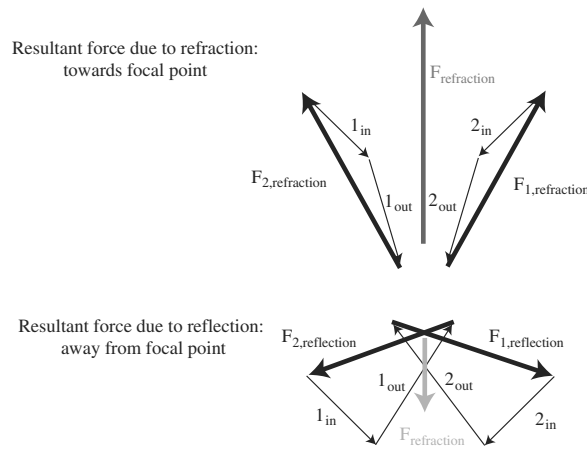
### **Optical Tweezers**

The forces exerted on small objects can be either repulsive, due to light reflection, or attractive, due to refraction (see Fig. 17.45). As the sphere modifies the direction of the momentum of incident photons, an equal and opposite momentum is transferred to the sphere with an associated force in the direction of the focal point of the beam focused on the sphere. This attracts the sphere towards the centre of this focal point. In contrast, the force due to reflection on the sphere tends to push it away from the focal point. To reduce the effect of this repulsive force, objectives with very large numerical aperture are used, in such a way as to focus the light as precisely as possible on the sphere and favour highly inclined rays in the trapping force. With an objective immersed in an oil of refractive index 1.5, numerical apertures of 1.4 are possible.

Two light rays 1 and 2 are refracted in a dielectric sphere immersed in water. The directions of these rays obey Snell's law  $n_{\text{water}} \sin \theta_{\text{water}} = n_{\text{sphere}} \sin \theta_{\text{sphere}}$ . Figure 17.45 illustrates the situation without taking into account multiple reflections inside the sphere.



**Fig. 17.45.** Forces on a small sphere due to refraction and reflection of a focused beam



**Fig. 17.46.** The change in the momentum of photons passing through the sphere leads to a net force

*Measuring Forces with Optical Tweezers*

The effect can be gauged quantitatively and the force on a sphere can actually be measured. This effect is particularly useful in biology. A sphere, typically with diameter between 100 nm and 100  $\mu\text{m}$ , positioned in a Gaussian beam focused by a microscope objective, will be affected by the steep gradients in the intensity profile of the beam in the three space directions, which tend to draw the sphere towards the centre of the focal point. The position of the

sphere can be localised using an imaging technique, e.g., projecting onto a quadrant photodiode, which can attain accuracies in the nanometric range.

If the sphere is displaced from the centre of the focal point, it is subject to a restoring force proportional to its distance from the focal centre:  $\mathbf{F} = -k\mathbf{r}$ , where  $\mathbf{r}$  is the vector from the sphere to the focal centre and  $k$  is the restoring coefficient. In practice, the sphere moves all the time due to Brownian motion. However, if it moves away from the focal point under the action of an external force exerted by a protein attached to it, for example, the force attracting it back can be quantified and will depend on the force exerted by the protein. By measuring the displacement of the sphere, the value of the force can be deduced with an accuracy that depends on the restoring coefficient, which itself depends on the trapping technique and the size of the sphere. Measured values of the restoring coefficient are generally of the order of 50 pN/ $\mu\text{m}$ , whilst the most sophisticated devices can reach 0.5 pN/ $\mu\text{m}$ .

An optical trapping experiment requires a microscope equipped with an objective with large numerical aperture. Trapping occurs under continuous illumination in the infrared, since this wavelength is the best suited to cause least damage to biological media like cell membranes. The optical trapping beam is directed onto the bead by a set of mirrors and lenses, whilst an acousto-optic modulator is used to control its direction.

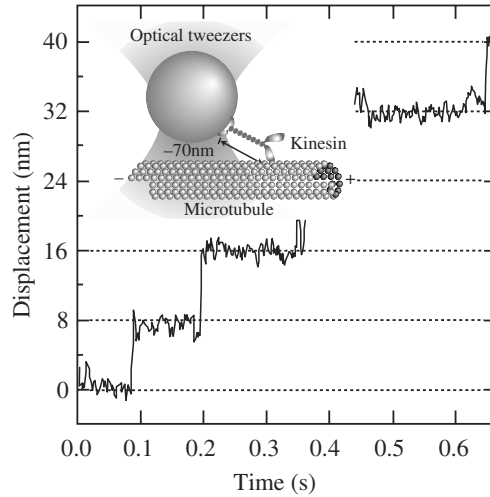
The instrument must be calibrated in order to be able to infer the force exerted on the sphere. To do so, a known force is applied and the displacement of the sphere is measured, which yields the value of the restoring coefficient  $k$  for the system. Such a force can be provided by the known flux of a fluid with viscosity  $\eta$ , whence  $\mathbf{F}_{\text{vis}} = 6\pi\eta a\nu\mathbf{u}$ , where  $\nu$  is the fluid velocity, and  $a$  the sphere radius. This force can be modulated in time for greater accuracy. Other calibration techniques are based on position fluctuations due to Brownian diffusion, and in particular, their frequency spectrum.

Optical tweezers are no used in many applications. Two examples are shown in Figs. 17.47 and 17.48. The first case concerns the displacement of molecular motors (kinesin protein) isolated on a microtubule. The second monitors the DNA transcription process, the role of the molecular motor being played by RNA.

### Microscopy Techniques Using Tips

The techniques discussed here are treated in more detail in Chaps. 3–5. We therefore restrict the following to features relevant to biological problems.

After twenty years of research in this field, the techniques of probe microscopy are still in the development stage, particularly with regard to image interpretation [64]. This type of technique is especially relevant to the study of surfaces and the manipulation of atoms and molecules. The basic setup of the scanning tunneling microscope, atomic force microscope, or near-field optical microscope is a scanning system based on piezoelectric transducers, a feedback loop, and a system for recording data. Local measurements are then



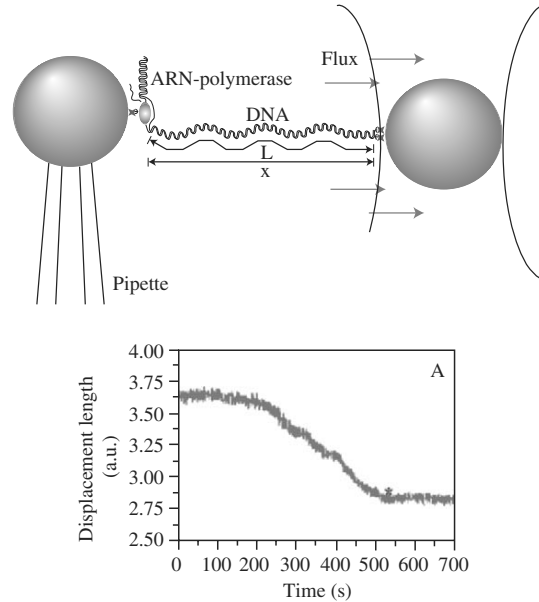
**Fig. 17.47.** Molecular motors and optical tweezers. The example here is kinesin, a small protein in the form of a dimer, capable of moving along a microtubule. Attaching a single kinesin protein to a polystyrene bead trapped by optical tweezers, its motion can be revealed. It occurs in discrete steps of 8 nm [61]. Every 8 nm, the protein recovers energy by hydrolysing an ATP molecule to form ADP [62]

possible by establishing the feedback conditions. For the observation of single molecules, this type of approach has opened up a wide range of channels for exploring their physical, chemical, mechanical and electronic properties. The type of molecule concerned varies from diatomic systems to complex biological entities. Moreover, micro- and nanotechnologies are now capable of fabricating accurate probes on scales that approach those of the molecular systems they investigate.

#### *Scanning Tunneling Microscopy (STM)*

The underlying idea of the scanning tunneling microscope is to cause a current to pass by the tunnel effect through the space separating the tip and the sample surface (see also Chap. 3). The feedback system holds the tunneling current constant. In the vacuum, this current typically changes by an order of magnitude for each 1-Å variation in the tip-sample separation. The voltage applied to the piezoelectric transducer in order to hold this current constant is then recorded, providing highly accurate topographical information concerning the sample surface.

The energy of the tunneling electrons varies under changes in the applied bias voltage between the tip and surface, and depending on the the sign of this bias, filled or empty electronic states can be probed. Observing molecules adsorbed on the surface, it is thus possible to correlate the surface topography with direct chemical interpretations related to the observed electronic



**Fig. 17.48.** Transcription and optical tweezers. A transcription complex (RNA polymerase, DNA) is attached between two spheres and held under a continuous flux. When a polymerase moves along the DNA, it tries to pull the two spheres along with it. One of the beads is retained by optical tweezers and the other by a micropipette. The measured displacement gives the distance between the two ends of the DNA and the transcription rate can be determined from it [63]

structures, since the bias can be used to select specific states. This technique has long been applied to semiconductor nanostructures, for example.

#### *Atomic Force Microscopy (AFM)*

In atomic force microscopy, the servo-system acts on the force as a function of the separation between tip and sample (see also Chap. 4). An AFM can work with or without contact between the tip and sample. In contact mode, one exploits the repulsive force between the probe and the surface, which varies rapidly as a function of the tip-sample separation. This mode is rarely used for biological samples, owing to their fragility. An alternative is the tapping mode, which attempts to minimise contact and causes less damage to samples. The feedback system is nevertheless still based on the contact time. In non-contact mode, AFM uses a weaker attractive force between the tip and sample. In each case, the recorded signal comprises measurements of the deformation of the tip holder, e.g., by deflection of a light signal reflected from the latter during its motion. It is also possible to measure the lateral motions of the tip in order to examine the friction and adhesion effects it undergoes.



AFM has been used intensively in surface topography studies, and more recently, in the manipulation of biological systems in aqueous media [65]. This method can be used to obtain information concerning the structure (elasticity, mechanical properties) of such systems on the molecular level.

## 17.4 Conclusion

As this discussion is intended as an introduction, the aim has not been to cover the whole field of nanobiophotonics, although the authors hope that this neologism will appear a great deal clearer to the reader on reaching the present section. Nor has it been our desire to represent these issues in the rather rigid way one might expect a mature field, when the subject is still in its infancy. At best, we will have identified the main lines of force as they stand today, and the reader is warned not to extrapolate too willingly beyond the most immediate future, when we are only just dealing with the definition stage.

As a conclusion then, we shall simply recall here some of the applied and cognitive matters raised by current and future progress in nanobiophotonics, without risking a judgement in the heated debate that surrounds the chances of success for one approach or another.

Concerned as always with the nanoscale, one of the major alternatives today which will certainly continue to develop over the next few years is exogenous photonic marking, as opposed to the endogenous response of biological media. In the first case, nanoparticles of all types, endowed with diverse and identifiable physical properties, are adopted as accessories, provided that they insinuate themselves into the medium as discretely as possible, whilst providing the microscopic device with the required luminosity and spatial resolution associated with the properties of nanostructures that have been optimised with this in mind. It goes without saying that the so-called endogenous response of the tissue alone is in every way preferable, particularly for *in vivo* studies, given the drawbacks of marking by a degradable particle that is foreign to the biological ecosystem, although at the same time clearly falling short of the record performance that can be achieved by introducing a nanostructure. Since each approach has its advantages and disadvantages, they will doubtless be called upon to complement and emulate one another, rather than just to compete, over the coming years.

Research into exogenous labels is itself far from finalised and still gives rise to lively controversy. Although semiconductors in the form of quantum dots based on III–V or II–VI alloys tend to be predominant in biophotonics, these also display some disadvantages and one may predict the emergence of still conjectural alternatives. An example is perhaps the metallic nanostructures which, when heated by laser flux, can provide an excellent source of information on this scale, without the risk of bleaching or blinking, playing upon the sensitivity of interferometry as a means of detection. Along the same lines, we

may also mention luminescent rare earth nanostructures functioning in the infrared, which allows their use deep inside tissues, or nanostructures made from frequency-doubling nonlinear molecular crystals, which have a very efficient response due to the coherent nature of the effect and which, in contrast with one- or multiphoton luminescence effects, do not involve resonances likely to cause photodegradation.

More generally, as far as instrumentation is concerned, one may expect a generalisation and growth of methods which combine advanced microscopy, especially confocal microscopy, with interferometric and polarimetric devices, rich sources of otherwise inaccessible information. As regards techniques using nonlinear ellipsometry, one may envisage local mapping of the generalised anisotropies of biological media, capable of reflecting nanofluxes of chemical species, or discontinuities or breaks in the electric field. At the same time, determination by means of microscopic interferometry of the contribution of the phase in the local response of biological systems should be able to refine otherwise incomplete structural information gleaned from intensity measurements alone, recalling that phase data is ignored in the incoherent effects commonly solicited, such as fluorescence.

Moreover, the coherent or incoherent nonlinear effects associated with the intrinsic response of tissues or of nanolabels often turn out to be highly sensitive to any kind of external perturbation affecting physicochemical, electric, or structural parameters describing the environment and operating at scales from the nanoscale to the mesoscale.

We are just beginning to be able to steer nano-objects towards designated targets such as lesions or tumors of a certain type by adorning their periphery with systems of monoclonal antibodies specific to these targets. Not only are novel nanodiagnostic techniques in view, but one may now glimpse the possibility of nano-intervention, using an extra drug release layer in the nanostructure which can be photo-triggered, as happens in current methods of dynamic phototherapy for cutaneous or subepidermal pathologies.

On a more fundamental level, different forms of radiation are likely to play an increasing part in guiding or at least assisting the traffic of reporter nanoparticles or biochemical species, such as repair genes carried by deactivated viruses but which still struggle to get past the various membranes which oppose their penetration to reach the core of the nucleus in sufficient numbers. From this point of view, the authors of this chapter are particularly attentive to the possibilities of applying coherent dual-frequency control techniques to biological systems, e.g., interference of multiphoton absorption paths, the simplest combining one- and two-photon absorptions. These allow one to orient and direct at will the motion of certain well-designed molecular systems, e.g., photo-isomerisable systems, through solid environments and in a potentially more precise way than with current optical tweezers in vacuum.

Finally, emulating the revolution in optoelectronics over the last two decades, where fundamental and applied research have moved forward hand

in hand,<sup>1</sup> fundamental repercussions are expected in biophotonics from research in biotechnology, especially from the spectacular development of DNA biochips over the last ten years. Capitalising on progress in the technology of silicon components, these DNA chips are beginning to provide fundamental research of the post-genome era with tailor-made multiple receptacles of a combinatorial nature which can be interrogated and analysed in real time by ultrahigh resolution read techniques associated with ever more powerful image analysis methods. With little risk of error, one may predict that new types of optoelectronic component, similar to those currently being developed for information technology, will emulate the development of new generations of DNA chips provided with internal photonic functionalities, and this all the more easily in that some are already based on the implementation of polymers and functionalised molecules, thereby well-placed to cooperate with biological systems.

Apart from the main pathologies (HIV, cancer), another wide field of applications is in neurophysiology. This is undoubtedly one of the new frontiers which remains the most open and the most fascinating for biophotonics. The stakes are twofold. One goal is to open observational windows that can combine the various levels of cooperation, in a way that was previously inaccessible, from the systemic (cortex, neural networks) to the subcellular (synapses and the interior of the neurone). There is no doubt that this will contribute to the advance of the cognitive sciences. The other goal is to bring to bear the resources of photonics on the problem of the neurodegenerative pathologies, which now stand as the next major challenge to be faced by our society.

We hope to have convinced the reader that the union between photonics and the life sciences is not some ephemeral consequence of a chance encounter. Indeed, upstream of the spectacular progress which is beginning to renew and transform the methodology and equipment of research laboratories in the life sciences, and which will very likely soon enter the clinical field, the basic problem situation which currently associates photonics with the life sciences is in fact of precisely the same kind as the problematic which already associated the nascent science of biology with the emergence of the first microscopes at the beginning of the seventeenth century, and which has never ceased to unite them. For when all is said and done, to see is to understand.

---

<sup>1</sup> As an example, recall that the observation of the fractional quantum Hall effect at the end of the 1980s arose directly from the discovery in the previous decade of techniques for depositing thin layers with accuracies that could be controlled on the atomic scale, which opened the way to the epitaxial growth of semiconductor interfaces, at a time when the search was on for ultrafast data processing techniques.

## References

### Introduction and Section One

1. Charra, F., Agranovitch, V.M., Kajzar, F. (Eds.): *Organic Nanophotonics*, Vol. 100, Nato Science Series II. Mathematics, Physics and Chemistry, Kluwer Academic Publishers, Dordrecht (2003)
2. Born, M., and Wolf, E.: *Principles of Optics*, MacMillan, New York (1964)
3. Saleh, B.E.A., and Teich, M.C.: *Fundamentals of Photonics*, Wiley, New York (1991)
4. Betzig, E., and Chichester, R.J.: *Science* **262**, 1422 (1993)
5. Tan, W., and Kopelman, R.: Subwavelength molecular exciton probes. In: *Molecular Electronics*, ed. by J. Jortner and M. Rattner, Blackwell Science, Oxford (1997)
6. J. Zyss (Ed.): *Photonique Moléculaire: Matériaux, Physique et Composants*, C.R. Physique **3**, No. 4 (2002)
7. Weiss, S.: *Science* **283**, 1676–1683 (1999)
8. Bachand, G.D., et al.: *Nano Letters* **1** (1), 42–44 (2000)
9. Nealey, P.F., Black, A.J., Wilbur, J.L., and Whitesides, G.M.: Micro- and nanofabrication techniques based on self-assembled monolayers. In: *Molecular Electronics*, ed. by J. Jortner and M. Rattner, Blackwell Science, Oxford (1997)
10. Steinmeyer, G.: *J. Opt. A: Pure Appl. Opt.* **5**, R1–R15 (2003)
11. Chemla, D.S., and Zyss, J. (Eds.): *Nonlinear Optical Properties of Organic Molecules and Crystals* (2 vols.), Academic Press, Orlando (1987)
12. Tamarat, Ph., Maali, A., Lounis, B., and Orrit, M.: *J. Phys. Chem.* **104** (1), 1–16 (2000)
13. Sunney Xie, X., and Trautman, J.K.: *Ann. Rev. Phys. Chem.* **49**, 441–480 (1998)
14. Boulanger, B., and Zyss, J.: Nonlinear optical properties. In: *International Tables for Crystallography*, Chap. 17, Vol. D (*Physical Properties of Crystals*), ed. by A. Authier, Kluwer, Dordrecht (2003) pp. 178–219

### Section Two

#### Fluorescence

15. Lakowicz, J.R. (Ed.): *Principles of Fluorescence Spectroscopy*, Kluwer, Dordrecht (1999)

#### Molecular Engineering

16. Zyss, J., Ledoux, I., Nicoud, J.F.: In: *Advances in Molecular Engineering for Quadratic Nonlinear Optics: Materials, Physics and Devices*, ed. by J. Zyss, Academic Press, New York (1994)
17. Albota, M., Beijonne, D., Brédas, J.-L., Ehrlich, J.E., Fu, J.-Y., Heikai, A.A., Hess, S.E., Kogej, T., Levin, M.D., Marder, S.R., McCord-Maughon, D., Ferry, J.W., Röckel, H., Rumi, M., Subramaniam, G., Webb, W.W., Wu, X.-L., and Xu, C.: *Science* **281**, 1653–1656 (1998)

## Nanoparticles

18. Boyer, D., Tamarat, P., Maali, A., Lounis, B., Orrit, M., *Science* **297**, 1160–1163 (2002)
19. Bruchez, M., Moronne, M., Gin, P., Weiss, S., Alivisatos, A.P.: *Science* **281**, 2013–2016 (1998)
20. Chan, W.C.W., Nie, S.: *Science* **281**, 2016–2019 (1998)
21. Marks, R.S., et al.: *Mat. Sci. Eng. C* **21**, 189–194 (2002)
22. Niemeyer, C.M.: *Angew. Chem. Int. Ed.* **40**, 4128 (2001)
23. Valeur, B.: *Molecular Fluorescence: Principles and Applications*, Wiley-VCH (2002)
24. de Silva, A.P., et al.: *Trends in Biotechnology* **19** (1), 27 (2001)

## Section Three

### Single-Molecule Detection by Fluorescence

25. Bashé, T., Moemer, W.E., Orrit, M., Wild, U.A. (Eds.): *Single-Molecule Detection, Imaging and Spectroscopy*, Verlag Chemie (1997)
26. Moemer, W.E., Orrit, M.: Illuminating single molecules in condensed matter, *Science* **283**, 1670–1676 (1999)
27. Duchesne, L., et al.: *Science* **292**, 255–258 (2001)
28. Tamarat, P., Maali, A., Lounis, B., Orrit, M.: *J. Phys. Chem. B*, **104** (1), 1–16 (2000)
29. Ambrose, W.P., et al.: *J. Chem. Phys.* **95**, 7150 (1991)
30. Xie, X.S., Trautman, J.K.: *Aimu. Rev. Phys. Chem.* **49**, 441–480 (1998)
31. Ishii, Y., Yanagida, T.: *Single Mol.* **1** (1), 5–13 (2000)
32. Rigler, R., Orrit, M., Bashé, T. (Eds.): *Single-Molecule Spectroscopy*, Nobel Conference Lectures, Springer Series in Chemical Physics, Vol. 67 (2001)
33. Michalet, X., Weiss, S.: *C. R. Physique* **3**, 619–644 (2002)
34. Pawley, J.B.: *Handbook of Biological Confocal Microscopy: Foundations of Confocal Scanned Imaging in Light Microscopy*, 2nd edn., Plenum Press, New York (1995)
35. Pierce, D.W., et al.: *Nature* **388**, 338 (1997)
36. Schaffer, J., Volkmer, A., Eggeling, C., Subramaniam, V., Striker, G., Seidel, C.A.M.: Identification of single molecules in aqueous solution by time-resolved fluorescence anisotropy, *J. Chem. Phys. A* **103** (3), 332–336 (1999)
37. Schwille, P., Haustein, E.: Fluorescence Correlation Spectroscopy. A tutorial for the *Biophysics Textbook Online* (BTOL), [www.biophysics.org/](http://www.biophysics.org/) (2002)
38. Cagnet, L., Coussen, F., Choquet, D., Lounis, B.: *C. R. Physique* **3**, 645–656 (2002)
39. Lu, H.P., et al.: *Science* **282**, 1877 (1998)
40. Cantor, C.R., Schimmel, P.R. (Eds.): *Biophysical Chemistry*, Part II: Techniques for the study of biological structure and function, W.H. Freeman, New York (1980)
41. Deniz, A., et al.: *Proc. Natl. Acad. Sci.* **97**, 5179 (2000)

42. From the Dawson college website: <http://omega.dawsoncollege.qc.ca>
43. Schmidt, T., Schütz, G.J., Baumgartner, W., Gruber, H.J.: Proc. Natl. Acad. Sci. USA **93**, 2926–2929 (1996)
44. Peters, I.M., de Grooth, B.G., Schins, J.M., Figdor, C.G., Grève, J.: Rev. Sci. Instr. **69** (7), 2762–2766 (1998)
45. Urlic, M., et al.: J. Biophys. **83**, 2681–2692 (2002)
46. Schütz, G.J., et al.: EMBO Journ. **19** (5), 892 (2000)
47. Seisenberg, G., et al.: Science **294**, 1929 (2001)

### Multiphoton Microscopy

48. Centoze, V.E., et al.: Biophys. J. **75**, 2015–2024 (1998)
49. Lakowicz, J.R., (Ed.): *Topics in Fluorescence Spectroscopy: Nonlinear and Two-Photon-Induced Fluorescence*, Vol. 5, Kluwer, Dordrecht (1997)
50. Bagatolli, L.A., et al.: Biophysical J. **77**, 2090 (1999) and associated website
51. Denk, W., Strickler, J.H., Webb, W.W.: Science **248**, 73 (1990)
52. Lagurné-Labarthe, F., and Shen, R.: In: *Optical Imaging and Microscopy: Techniques and Advanced Systems*, ed. by F.J. Kao and P. Torok, Springer Series in Optical Sciences, Vol. 87, Springer Verlag (2003)
53. Bozhevoinyi, S.L, Geilser, T.: J. Opt. Soc. Am. A **15** (8), 2156–2162 (1998)
54. Moreaux, L., et al.: Opt. Lett. **25**, 320 (2000)
55. Silberberg, Y., et al.: Optics Express **5** (8), 169 (1999)
56. Cheng, J.X., Jia, Y.K., Zheng, G., Xie, S.: Biophys. J. **82**, 502–509 (2002)
57. Zumbusch, A., et al.: Phys. Rev. Lett. **82**, 4142 (1999)

### Optical Tweezers

58. Ashkin, A., et al.: Opt. Lett. **11** (5), 288–290 (1986)
59. Mehta, A.D., Spudich, J.A., Smith, D.A., Simmons, R.M.: Science **283**, 1689–1695 (1999)
60. Bustamante, C., Macosko, J.C., Wuite, G.J.L.: Nature Reviews **1**, 130–136 (2000)
61. Svoboda, K., et al.: Nature **365**, 721 (1993)
62. Schnitzer, M.J., et al.: Nature **388**, 6640 (1997)
63. Davenport, R.J., et al.: Science **287**, 2497 (2000)

### Techniques Using Tips

64. McCarthy, O.S., Weiss, P.S.: Chem. Rev. **99**, 1983–1990 (1999)
65. Liphardt, J., Onoa, B., Smith, S.B., Tinoco, I., Bustamante, C.: Science **292**, 733–738 (2001)

## Numerical Simulation

X. Blase and C. Delerue

It is not easy to investigate the structural and electronic properties of the systems used in nanotechnology, simply because they are so small. Despite considerable progress made with near-field techniques, observation of matter with nanometric resolution is still a challenge.

In order to obtain a better understanding of materials on the atomic scale, experiment has found itself a useful ally in the form of numerical simulation. This branch of physics, standing midway between theory and experiment, is well known to the general public in the context of hydrodynamics. Indeed, weather prediction or aerodynamic modelling of a prototype car or plane can be carried out on the computer, without the need for wind tunnel tests. This same approach, where one seeks to reproduce the behaviour of matter on a given scale in the computer, has come a long way in materials science since its beginnings at the end of the second world war, and is now accompanying the development of nanoscience. The idea here is to simulate the behaviour of matter on the atomic scale.

The aim of the present chapter is not to reproduce the theory of structural, electronic, magnetic, optical or transport properties of nanostructures. These subjects and the relevant fundamental equations have been discussed in earlier chapters. Rather, we shall show how to obtain the ingredients required to implement these theories for a specific real system. In particular, we shall show that numerical simulations can be used to determine atomic and electronic structures, i.e., wave functions and energy levels, of a given material. This characterisation of the atomic and electronic state provides a complete description of the system, whereby its intrinsic features and its interaction with the environment (STM tip, electromagnetic field, etc.) can be understood. Other examples will also be mentioned.

## 18.1 Structural Properties

In order to study the properties of a material using numerical simulations, its atomic structure must first be ascertained. Indeed, the relative position of the atoms in space, the type of chemical bonds, and the level of phase segregation in the case of systems comprising several types of atom are all factors with a strong influence on the structural properties, i.e., elasticity, hardness, melting temperature, etc., and the electronic properties, i.e., band structure, optical absorption, magnetism, etc., of the ensemble.

One might be tempted to think that nanoscale systems such as clusters will behave like a blob of matter extracted from the bulk material. However, this intuitive approach often proves unsatisfactory. Even in the case of relatively large structures, e.g., clusters of large radius, although the core of the object may retain a structure close to that of the bulk solid, the surface atoms generally occupy very different positions. Indeed, compared with their environment in the solid, these atoms have lost neighbours. Chemical bonds have been broken. To make up for this reduction in coordination number, surface atoms will shift to create new bonds or stronger bonds with neighbouring atoms. This phenomenon is known as surface reconstruction. Among other things, it has a significant effect on the chemical reactivity of the cluster. The same type of phenomenon occurs in clusters imbedded in a matrix. The chemical environment, which differs from that in the bulk solid as far as the surface atoms are concerned, leads to varying degrees of structural rearrangement.

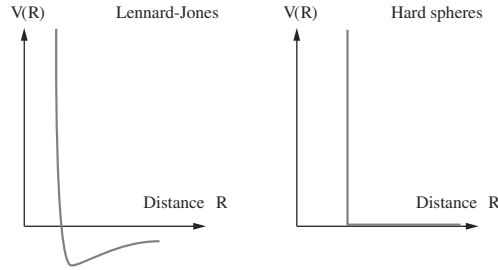
In the case of relatively small structures (bearing in mind that, in a cube containing 1 000 atoms, about 60% are at the surface!), any cluster can reconstruct to minimise its energy, i.e., to move towards greater stability. To be able to assess the stability of a structure, one must have methods capable of calculating the interaction potential energy between atoms. As we shall see below, one can then also calculate the forces those atoms will exert on each other.

### 18.1.1 Interatomic Potentials and Forces

The first approaches to calculating the energy of an atomic system are based on the use of empirical potentials capable of reproducing as faithfully as possible the interaction potential energy between two atoms a distance  $R$  apart. Intuitively, we know that two atoms repel one another at close range and do not interact at all at long range. Their interaction energy goes through a minimum corresponding to the equilibrium interatomic distance. This behaviour is represented by the potential profile shown in Fig. 18.1.

The empirical approach consists in postulating a parametrised functional form for the distance dependence of the potential energy. In his pioneering work in 1922, the English physicist Lennard-Jones thus proposed the relationship





**Fig. 18.1.** (a) Lennard-Jones potential. (b) Hard-sphere potential

$$V(R) = A \left[ \left( \frac{\sigma}{R} \right)^{12} - \left( \frac{\sigma}{R} \right)^6 \right] \quad (18.1)$$

to describe the interaction between two atoms. The terms in  $1/R^{12}$  and  $1/R^6$  are called the repulsive and attractive terms, respectively. The parameters  $A$  and  $\sigma$  are fitted to reproduce certain properties of the system one hopes to describe, such as the equilibrium separation, vibrational frequencies, compressibility, etc. This is indeed an empirical approach, since one must have a priori knowledge of the material under investigation. The parameters depend not only on the chemical species present, but also on the crystal structure.

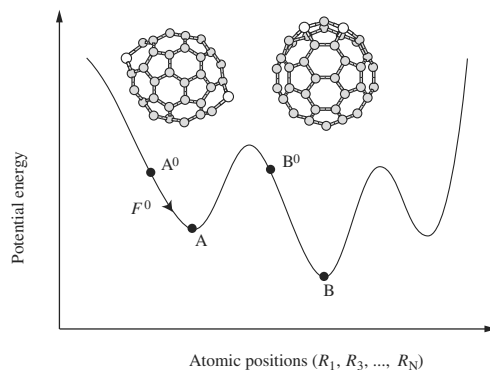
The empirical approach has been further developed and is still widely used. Other functional forms have since been suggested. Depending on whether the system is ionic, metallic, or covalent, the form of  $V(R)$  can change significantly (see Chap. 7 and in particular Sect. 7.2.1). In each case,  $V(R)$  can be used to calculate the total potential energy  $E^{\text{pot}}$  of the system by summing over all pairs:

$$E^{\text{pot}}(\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N) = \frac{1}{2} \sum_{I \neq J} V(|\mathbf{R}_I - \mathbf{R}_J|), \quad (18.2)$$

where  $(\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N)$  are the positions of the  $N$  atoms making up the system. In principle, to find the structure with lowest energy, one merely has to minimise the energy  $E^{\text{pot}}$  as a function of the  $3N$  variables  $(\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N)$ . We shall not go into the details of the mathematical methods available for minimising multivariable functionals. From the point of view of the physicist, let us just note that the total force exerted on an atom located at  $\mathbf{R}_I$  can be found from the total energy as

$$\mathbf{F}_I = - \frac{d}{d\mathbf{R}_I} E^{\text{pot}}(\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N). \quad (18.3)$$

Calculation of forces allows one to carry out molecular dynamics simulations (see Sect. 18.1.3), but also to minimise the energy of the system starting from a given geometry, by displacing the atoms in the direction of the force exerted on them:  $\mathbf{R}_I^{n+1} = \mathbf{R}_I^n + \lambda \mathbf{F}_I$ , where  $\mathbf{R}_I^{n+1}$  and  $\mathbf{R}_I^n$  are the positions of



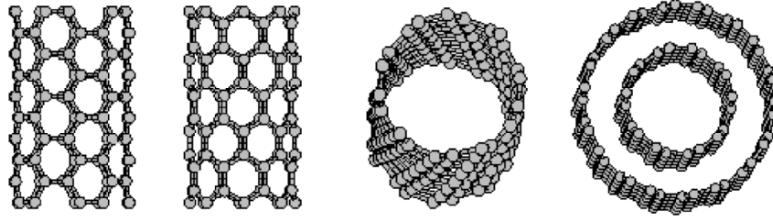
**Fig. 18.2.** Schematic potential energy surface for a fullerene doped by substituting two silicon atoms. Two possible isomers are shown, differing in the relative position of the silicon atoms. Each one corresponds to a different energy minimum

atom  $I$  in the  $(n + 1)$ th and  $n$ th iterations of the minimisation process. The parameter  $\lambda$  is an adjustable positive parameter controlling the rate of convergence. The first order relation  $dE^{\text{pot}} = -\mathbf{F}_I d\mathbf{R}_I = -\lambda \mathbf{F}_I^2$  shows that this algorithm does indeed lead to a minimisation of the energy. At equilibrium, as the forces are then zero, the atoms will move no longer and the process converges. This method, known as the method of steepest descent, is a robust one and widely used. Other, more sophisticated techniques, e.g., conjugate gradients, Newton–Raphson method, derive from it.

### 18.1.2 Potential Energy Surface

The technique proposed above for finding the equilibrium structure of an atom ensemble conceals a crucial problem: the final solution may depend heavily on the structure used to initialise relaxation. To understand how this comes about, Fig. 18.2 shows a schematic graph of a possible function  $E^{\text{pot}}(\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N)$ . It is clear that this curve in the  $3N$ -dimensional space of atomic positions has a great many minima. If the system is prepared in state  $\mathbf{A}^0$  and the atoms are pushed in the direction of the forces exerted upon them, the system will move towards the minimum denoted  $\mathbf{A}$ , which differs from the one obtained by relaxing from the initial system  $\mathbf{B}^0$ . The minimum  $\mathbf{A}$  is called a local minimum, because it has greater energy than the system at  $\mathbf{B}$ . Determination of the absolute minimum in a space of  $3N$  variables is generally a very challenging problem. There are many available algorithms and it is difficult today to speak of a general method that is preferred over all the others.

Below we shall describe the techniques known as molecular dynamics, whose scope goes well beyond the determination of the minimum energy. To motivate this approach, note that nature does not necessarily choose the most



**Fig. 18.3.** Carbon nanotubes. *From left to right:* Single-walled armchair, zigzag, and chiral nanotubes, and multiwalled nanotube. The method of synthesis employed leads to one or other of these structures, depending on physical conditions such as temperature, the use of a catalyst, and so on

stable structure. Depending on the conditions of synthesis, e.g., temperature, pressure, presence of a catalyst, etc., one isomer or another, or a mixture of them, can be obtained. It is thus important to study matter in temperature and pressure conditions equivalent to those in the experiment. This is the aim of molecular dynamics, whose principle is described below.

### 18.1.3 Classical Molecular Dynamics

Molecular dynamics [1] is a simulation tool for following the trajectories of a group of atoms with positions  $\mathbf{R}_I(t)$ . The atoms are treated classically and obey Newton's law

$$M_I \frac{d^2 \mathbf{R}_I}{dt^2} = \mathbf{F}_I, \quad (18.4)$$

where  $\mathbf{F}_I$  is the force exerted on the atom indexed by  $I$ .

For an isolated ensemble of atoms, e.g., free clusters, the trajectory is fully defined by the initial conditions (positions and velocities) and Newton's law. Therefore, to follow the dynamic evolution of the system, one must know:

- the initial conditions,
- an integration algorithm,
- the forces  $\mathbf{F}_I$  on the atoms in a given configuration.

### Integration Algorithms

Numerical solution of the equations of motion is achieved by discretising the time parameter in Newton's equation. A widely used algorithm is the so-called Verlet algorithm. It is based on the truncated expansions

$$\begin{aligned} \mathbf{R}_I(t+h) &= \mathbf{R}_I(t) + h \frac{d\mathbf{R}_I}{dt} + \frac{h^2}{2} \frac{d^2\mathbf{R}_I}{dt^2} + \frac{h^3}{6} \frac{d^3\mathbf{R}_I}{dt^3} + O(h^4), \\ \mathbf{R}_I(t-h) &= \mathbf{R}_I(t) - h \frac{d\mathbf{R}_I}{dt} + \frac{h^2}{2} \frac{d^2\mathbf{R}_I}{dt^2} - \frac{h^3}{6} \frac{d^3\mathbf{R}_I}{dt^3} + O(h^4), \end{aligned}$$

which combine to give (with  $M_I = 1$ )

$$\mathbf{R}_I(t+h) = 2\mathbf{R}_I(t) - \mathbf{R}_I(t-h) + h^2\mathbf{F}_I(t) + O(h^4). \quad (18.5)$$

Hence, knowing  $\mathbf{R}_I(t)$ ,  $\mathbf{R}_I(t-h)$  and  $\mathbf{F}_I(t)$ , the position of the particle at a later time  $t+h$  can be ascertained. Moving forward in this way, the whole trajectory can be obtained.  $h$  is known as the time step in the numerical integration. From  $\mathbf{R}_I(t+h)$  and  $\mathbf{R}_I(t-h)$ , one can also calculate the velocity of the atoms by

$$\mathbf{V}_I(t) = \frac{\mathbf{R}_I(t+h) - \mathbf{R}_I(t-h)}{2h} - \frac{h}{12} [\mathbf{F}_I(t+h) - \mathbf{F}_I(t-h)] + O(h^3). \quad (18.6)$$

The term  $O(h^4)$  in (18.5) indicates that the position  $\mathbf{R}_I(t+h)$  is determined up to an error of the order of  $h^4$ . To obtain a high level of accuracy, i.e., to be sure that the discretised trajectory is always close to the true trajectory,  $h$  must be small in some well-defined way. The quality of integration can be adjusted by controlling the integrals of motion. In the case of an isolated system, i.e., a microcanonical ensemble, the total energy of the system, viz.,

$$E(t = nh) = \sum_I \frac{1}{2} M_I \mathbf{V}_I^2(t) + \frac{1}{2} \sum_{I \neq J} V(\mathbf{R}_I(t) - \mathbf{R}_J(t)), \quad (18.7)$$

must be independent of time. Here  $n$  indexes the time step and  $V(\mathbf{R}_I - \mathbf{R}_J)$  is the interaction potential for the two atoms at positions  $\mathbf{R}_I$  and  $\mathbf{R}_J$ .

### Exercise

Write a molecular dynamics program for one particle moving in one dimension in the case of a simple harmonic oscillator with equation of motion  $M d^2x/dt^2 = -kx$ . By choosing different initial conditions (position and speed), check that the numerical trajectory is close to the analytic solution provided that the time step  $h$  is small compared with the natural period of the oscillator. Study the evolution of the total energy of the system as a function of time.

### Temperature and Thermostats

As indicated above, the atomic structure of a material can depend sensitively on the conditions of temperature and pressure prevailing during its synthesis. Moreover, many physical processes such as melting, coalescence of two clusters, or chemical reactions between molecules are thermally activated. It is thus important in molecular dynamics simulations to be able to introduce the interaction of the system under investigation with its surroundings, e.g., surface, carrier gas, laser, since these determine the level of heat exchange.

Such exchange processes between systems can be complex. For example, it is difficult to describe the way a laser or an electrical current heats a material, and the temperature of a moving cluster or one deposited on a surface depends on the multiple collisions occurring with atoms in the carrier gas or at the surface, respectively (surface phonons, in the latter case). Rather than attempting to simulate all these complex processes in an exact manner, the approach adopted consists in bringing the system artificially to the desired temperature.

In order to make these things precise, let us consider the dynamics of a cluster, e.g., melting, fragmentation, etc. at a given temperature  $T$ . A first approach that is easy to apply consists in attributing velocities  $\mathbf{V}_I(t=0)$ , as initial conditions for the equations of motion, that have been prepared so to speak, so that their average value is in agreement with the heat energy:

$$\sum_{I=1,N} \frac{1}{2} M_I \mathbf{V}_I^2 = \frac{3}{2} N k_B T,$$

where  $k_B$  Boltzmann's constant. However, if the system is now allowed to evolve without interaction with the surroundings, its energy will be constant, rather than its temperature. There will be an exchange between the potential energy and the kinetic energy and the temperature will fluctuate.

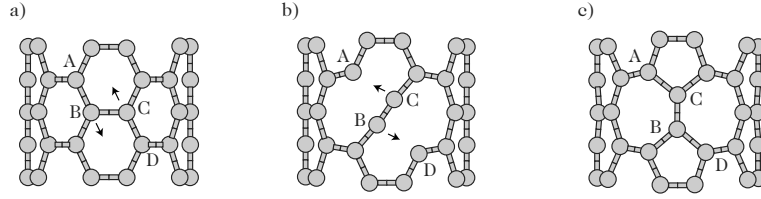
To remedy this problem and hold the temperature constant, a very simple technique is often employed, known as velocity renormalisation. At each time step, or after a fixed number of time steps, the velocities can be changed by a small factor  $\alpha$ , i.e.,  $\mathbf{V}_I(t) \rightarrow (1 + \alpha)\mathbf{V}_I(t)$ , in such a way as to keep the average velocity around  $3k_B T/2$ . Although rather crude and ad hoc, this technique gives good results. Not only is the temperature stabilised around the required value, but the energy distribution over all vibrational modes remains in reasonable agreement with the distribution predicted by statistical physics, i.e., the Boltzmann distribution.

Other more sophisticated techniques known as thermostat techniques have been proposed. In these approaches, there are exact results which can be used to show that the heat transfer towards the atomic system does indeed lead to a Boltzmann distribution. However, the simple approach discussed here remains widely used. Other thermostat techniques can also stabilise the pressure around a specified value.

To illustrate, Fig. 18.4 shows three snapshots of a molecular dynamics simulation for a carbon nanotube under tensile stress, revealing the so-called Stone–Wales transformation (formation of 5/7 cycles). This transformation, first ‘discovered’ through simulations, is useful for understanding the plastic properties, e.g., hardness, ductility, of carbon nanotubes.

#### 18.1.4 Monte Carlo Methods

Another large class of simulation methods which appeared at the same time as molecular dynamics, just after the second world war, go by the name of



**Fig. 18.4.** Stone–Wales transformation (formation of 5/7 cycles) in a carbon nanotube under tensile stress

the Monte Carlo methods. In contrast to molecular dynamics, the idea here is not to follow the trajectory of the system in time, but rather to sample the available configuration space in an efficient way. For concreteness, recall that for a dynamical system (like a system at finite temperature), any physical observable, e.g., band gap, magnetic moment, coordination number, etc., is a time average of the form

$$\langle M \rangle = \int_0^\tau M(t) \frac{dt}{\tau},$$

of the instantaneous values  $M(t)$  taken by the system for a given configuration  $[\mathbf{R}_1(t), \dots, \mathbf{R}_N(t)]$ . This average can be found using molecular dynamics.

Another approach is based on a result arising from statistical physics. For a system at constant temperature, for example, this average can also be written

$$\langle M \rangle = \frac{1}{Z} \int d\mathbf{R}_1 \dots \int d\mathbf{R}_N M(\mathbf{R}_1, \dots, \mathbf{R}_N) \exp \left[ -\beta E^{\text{pot}}(\mathbf{R}_1, \dots, \mathbf{R}_N) \right], \quad (18.8)$$

where the exponential term includes the Boltzmann factor ( $\beta = 1/k_B T$ ) and  $Z$  is a normalisation factor. This is therefore an ensemble average over all possible values ( $\mathbf{R}_I, I = 1, \dots, N$ ) rather than a time average. The ergodic theorem shows that these two averages must be equal in the long time limit  $\tau \rightarrow \infty$ .

Direct evaluation of the integral in (18.8) is difficult, however. Consider a system of 10 atoms, able to move around in a box of side 10 Å. Discretising the box by means of a mesh of interval 1 Å in each direction, we obtain  $10^{30}$  possible configurations if we allow the atoms to move between the nodes of the lattice. This number is far too big to be able to find the exact average over all possible configurations.

The Monte Carlo approaches use the fact that all the configurations with high energy  $E^{\text{pot}}(\mathbf{R}_1, \dots, \mathbf{R}_N)$  contribute little to the average owing to the presence of the Boltzmann exponential term. The main idea here is thus to keep only the lowest energy configurations. One speaks of selective phase space sampling techniques.

To implement this selective sampling, the following technique is used. One starts with the configuration  $(\mathbf{R}_I^0, I = 1, \dots, N)$  of energy  $E^0$ . Small

displacements ( $d\mathbf{R}_I$ ) are then chosen at random to obtain a neighbouring configuration ( $\mathbf{R}'_I$ ) of energy  $E'$ . The ratio

$$\frac{\exp(-\beta E')}{\exp(-\beta E^0)} = \exp\left[-\beta(E' - E^0)\right]$$

is then calculated. If this ratio is large enough in some well-defined sense, i.e., greater than a given fixed value, this means that the system is moving towards a region of phase space that will contribute significantly to the integral. This displacement is therefore accepted. If on the other hand the ratio is small, the system is therefore moving in the ‘wrong’ direction and the displacement is rejected. A new shift ( $d\mathbf{R}_I$ ) is then chosen at random and the procedure is repeated in this way so that the system can sample a large region of available ( $\mathbf{R}_I$ ). The name ‘Monte Carlo’ arises from the parallel with drawing lots in games of chance.

## 18.2 Electron Properties

In order to study the properties of electrons, one must turn to the principles of quantum mechanics which govern their behaviour. In contrast with the classical approaches presented above, where the atom was considered as a point object, electrons and ionic nuclei must be treated as separate particles interacting via the Coulomb potential. We begin by summarising some results from quantum mechanics, referring the reader to the standard textbooks for further detail, and in particular to [2], which treats the subject of elementary inorganic clusters.

### 18.2.1 Basic Results from Quantum Mechanics

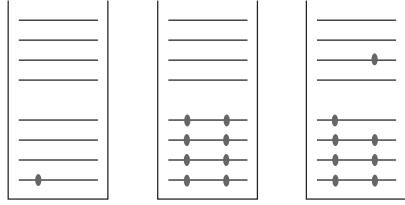
#### Independent Electrons

Whether one considers the well known example of an electron in a box, subject to the central potential of a proton in the hydrogen atom, or the periodic potential of the nuclei in a solid [3], the postulates of quantum mechanics [4] lead one to introduce wave functions  $\phi_n(\mathbf{r})$  that are solutions of the Schrödinger equation

$$\hat{H}\phi_n(\mathbf{r}) = \varepsilon_n\phi_n(\mathbf{r}),$$

where  $\hat{H}$  is the Hamiltonian of the system given by

$$\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r}). \quad (18.9)$$



**Fig. 18.5.** Confining potential and energy levels for an electron in a box (*left*) and several non-interacting electrons in the same box (*centre*). The figure on the *right* shows an excited state

where  $m$  is the electron mass,  $-\hbar^2\nabla^2/2m$  is the kinetic energy operator, and  $V(\mathbf{r})$  is the potential in which the electron is moving. The wave function  $\phi_n(\mathbf{r})$  gives the probability amplitude of finding the electron in the  $n$ th energy level  $\varepsilon_n$  at the point  $\mathbf{r}$  of space. When the system is not perturbed, the electron sits in the lowest energy level, known as the ground state (see Fig. 18.5, left).

Let us now consider a more complicated system, viz., a situation with several electrons. To a first approximation, it is common practice to treat the electron as non-interacting. This is the approximation used, for example, to provide a simple description of the Fermi gas, or the formation of energy bands in solids. In this approximation, each electron feels only the potential  $V(\mathbf{r})$ . Its wave function  $\phi_i$ , where the index  $i$  counts the electrons, thus satisfies the same Schrödinger equation as for the single-electron case. The set of available states ( $\varepsilon_n$ ) is the same as in the one-electron problem, and it is only the population of the levels that changes here. In the ground state, the levels are filled from the bottom, i.e., the lowest energies first, with a maximum of two electrons with opposite spins in each level. This is the Pauli exclusion principle for fermions. The energy difference between the highest occupied level and the lowest unoccupied level is called the band gap (see Fig. 18.5, centre). In solids, the occupied levels define the valence bands, whilst the empty levels define the conduction bands. During electron excitation, e.g., when the system interacts with an electromagnetic wave, an electron abandons an occupied level for an unoccupied one of higher energy, leaving a charge hole in the original level (see Fig. 18.5, right).

### General Case

The non-interacting electron model is an approximation. In reality, the electrons interact with one another via the Coulomb force between charged particles. The potential in which the electrons move thus depends on the whole set of positions ( $\mathbf{r}_i$ ,  $i = 1, \dots, N_e$ ). The Hamiltonian for the system is

$$\hat{H} = \sum_i^N \left( -\frac{\hbar^2}{2m} \nabla_i^2 \right) - \frac{e^2}{4\pi\epsilon_0} \sum_{I=1}^N \sum_{i=1}^{N_e} \frac{Z_I}{|\mathbf{r}_i - \mathbf{R}_I|} + \frac{e^2}{4\pi\epsilon_0} \sum_{i<j}^{N_e} \frac{1}{r_{ij}}, \quad (18.10)$$

where  $Z_I$  is the charge on the nucleus. Schrödinger's equation  $\hat{H}\phi = E\phi$  becomes an equation in  $3N_e$  variables, where  $N_e$  is the number of electrons.



This is called an  $N$ -body equation, and the Hamiltonian is an  $N$ -body Hamiltonian, in contrast to the one-body (or single variable) Hamiltonian given in the last section. The wave function solving this equation is a function of many variables,  $\phi(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_{N_e})$ , which describes the probability amplitude for having one electron at  $\mathbf{r}_1$ , another at  $\mathbf{r}_2$ , and so on. The exact solution of this differential equation to ascertain  $\phi$  gains rapidly in difficulty and becomes impossible when the number of electrons exceeds a few dozen. Likewise, even the calculation of the energy as the expected value of the Hamiltonian for the  $N$ -body wave function, viz.,  $\langle \phi | \hat{H} | \phi \rangle$ , becomes a multiple integral that is difficult to calculate in practice, and application of the variational principle<sup>1</sup> is no longer a feasible option. Approximations are thus required to treat the interaction between electrons.

In every case, the electron interaction is replaced by some sort of average interaction that does not depend explicitly on the positions of the other electrons. This so-called mean field is the same for all the electrons and converts the problem back to the simple case of non-interacting electrons. A first class of approximations is based on an empirical parametrisation of the electron–electron and/or electron–ion interactions, as was done for the interaction between atoms in the first section of this chapter, and can be done for the interaction between spins, which we shall return to briefly below.

A second family of methods aims to find approximations for the electron–electron potential on the basis of the principles of quantum mechanics, without appealing to empirical parameters. These are known as *ab initio* methods [5]. These approaches, which are generally more accurate but more time-consuming, will be discussed in Sect. 18.2.3. However, we begin with the semi-empirical approaches to electron structure.

### 18.2.2 Semi-Empirical Approaches to Electron Structure

To illustrate the relevance of the empirical methods, we first examine how they are applied to semiconductor nanostructures, e.g., composed of silicon. Consider a nanocrystal containing  $N$  atoms occupying positions  $\mathbf{R}_j$  ( $j = 1, \dots, N$ ). We assume that its electron structure can be described by a single-particle Hamiltonian  $\hat{H}$ . The idea is to find a certain number of eigenvalues  $\varepsilon_i^{e,h}$  and eigenstates  $\phi_i^{e,h}$  in the neighbourhood of the band gap:

$$\hat{H}\phi_i^{e,h} = \varepsilon_i^{e,h}\phi_i^{e,h}, \quad (18.11)$$

where the superscript e denotes empty states in the conduction band, whilst h indicates occupied states in the valence band.  $\hat{H}$  is the Hamiltonian of the electrically neutral system, but  $\varepsilon_i^e$  and  $\varepsilon_i^h$  are often interpreted as the energies

<sup>1</sup> According to the variational principle [4], the wave function  $\phi$  of the system in its ground state, i.e., at rest, minimises the energy  $E[\psi] = \langle \phi | \hat{H} | \phi \rangle$ . This is a very powerful way of obtaining the wave function  $\psi$  from the Hamiltonian  $\hat{H}$  using the standard mathematical algorithms for minimising functions (see Sect. 18.1.1).

of an extra electron and hole in the nanocrystal, which is an approximation.<sup>2</sup> As we have already seen, the Hamiltonian  $\hat{H}$  contains the kinetic energy of the electron, the potential energy of the interaction with the atomic nuclei, and the effective interaction potential with all other electrons.

The semi-empirical calculations assume that the Hamiltonian  $\hat{H}$  can be reasonably approximated by the Hamiltonian  $\hat{H}_0$  of the bulk solid within the nanostructure. This is the case, for example, in a nanocrystal of diameter  $d > 1$  nm. Semi-empirical methods such as the  $\mathbf{k} \cdot \mathbf{p}$  method and the effective mass method [7, 8], the pseudopotential method [9], and the tight-binding method [10, 11, 13] each propose a different approximation for  $\hat{H}_0$ . They involve a certain number of parameters which are fitted to the experimental data or the ab initio band structures. These parameters are then transferred to the nanostructures ( $\hat{H} = \hat{H}_0$ ) to which appropriate boundary conditions are applied, i.e., conditions describing the surfaces and interfaces. In general, only a small number of desired states  $\phi_i^e$  and  $\phi_i^h$  are directly calculated, which means that the semi-empirical methods can be used to study the electronic structure of much bigger nanostructures than ab initio methods. The quality of description of the band structure of the bulk solid and the relevance of the boundary conditions applied to the nanostructures are the two essential criteria whereby one may assess a semi-empirical method. We shall now describe several methods in more detail.

### Effective Mass Approximation

The simplest method for calculating the electron structure of semiconductor nanostructures is based on the effective mass approximation. To describe this, consider for example the case of an extra electron in a nanocrystal. (The crystal is then charged, but we shall neglect the induced Coulomb effect to simplify the problem.) When this electron is in the bulk semiconductor, it necessarily occupies a state at the bottom of the conduction band, since all the states of the valence band are occupied, by definition. The quantum theory of the electron structure then shows that this electron behaves effectively as a free electron, although with an inertial mass that is not the mass  $m_e$  of the electron, but rather an effective mass denoted  $m_e^*$  that is often smaller than  $m$ . In other words, when a force  $\mathbf{f}$  is applied to the electron, it has acceleration  $\gamma = \mathbf{f}/m_e^*$ , according to Newton's famous second law.

The effective mass reflects the fact that the particle is not moving in vacuum, but is surrounded by other electrons and nuclei with which it is continually interacting. The electron energy, and hence the Hamiltonian  $\hat{H}_0$ , are formally those of a free electron, thus reducing to the kinetic energy  $p^2/2m_e^*$ , where  $p$  is the magnitude of the momentum vector  $\mathbf{p}$  of the electron, with the zero energy fixed at the bottom of the conduction band ( $\varepsilon_c$ ). When the electron is in the nanocrystal, it interacts in approximately the same way with the

<sup>2</sup> The electric charges injected into a nanocrystal can be taken into account using corrections to the Hamiltonian derived from classical electrostatics.

other particles and it is reasonable to assume that its mass remains equal to  $m_e^*$ . The Hamiltonian for this confined electron can therefore be approximated by

$$\hat{H} = \hat{H}_0 + V_{\text{conf}}(\mathbf{r}) = -\frac{\hbar^2}{2m_e^*}\Delta + V_{\text{conf}}(\mathbf{r}), \quad (18.12)$$

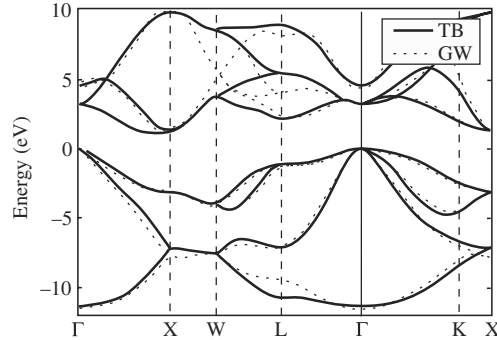
where  $V_{\text{conf}}$  is the confining potential simulating the effect of the surface. It reflects the fact that the electron prefers to remain within the nanocrystal and that the surfaces act as barriers for it.

A common approximation consists in setting  $V_{\text{conf}}(\mathbf{r}) = 0$  inside the nanocrystal and  $V_{\text{conf}}(\mathbf{r}) = V_0$  (constant) outside. In the limit  $V_0 \rightarrow +\infty$ , one obtains the equation for a free particle of mass  $m_e^*$  in an infinite potential well [8]. This problem has analytic solutions for a certain number of simple geometries, such as spherical and cubic wells. Generally, the eigenenergies  $\varepsilon_i^e$  of the confined electron vary as  $1/m_e^*d^2$ , where  $d$  is the characteristic size of the nanostructure, i.e., the radius for a sphere and the side for a cube. This is a well-known result for a particle enclosed in a box.<sup>3</sup> When  $d$  tends to infinity, the energy tends asymptotically towards  $\varepsilon_c$ , the energy at the bottom of the conduction band.

Since the situation is symmetric for electrons and holes (defining the mass of a hole as  $m_h^*$ ), we obtain the well-known result that the band gap of semiconductor nanostructures is broader than that of the bulk solid, and that this increase (the confinement energy) goes as  $1/d^2$ . We shall see shortly that this approximation is only valid in the weak confinement limit, i.e., the limit of small confinement energy or large size.

The level of difficulty involved in solving the Schrödinger equation with the Hamiltonian in (18.12) is independent of the number of atoms and the size of the system. This is a major advantage of the effective mass method and the  $\mathbf{k} \cdot \mathbf{p}$  methods derived from it. These methods successfully describe, among other things, the physics of 2D electron gases, e.g., quantum well lasers, MOS transistors, MESFET, etc., and quantum dots when they are not too strictly confined, e.g., stressed InAs/GaAs islands [14]. However, the effective mass approximation generally overestimates confinement energies in small semiconductor nanostructures because the description of the bulk solid Hamiltonian

<sup>3</sup> This variation is easily interpreted by the following qualitative argument. We have seen that the energy varies in the effective mass approximation according to  $p^2/2m_e^*$ . Moreover, in quantum mechanics, by the wave-particle duality, the momentum is related to the wavelength  $\lambda$  of the associated wave by the de Broglie relation  $p = h/\lambda$ . Finally, since the electron is confined, the electron wave must be stationary and must vanish on the nanocrystal surface, so that  $d$  is equal to a whole number multiple of  $\lambda$ . (The same relationship holds for the wavelengths associated with the vibrations of a string attached at each end and the length of the string.) Combining all these observations, we arrive at the fact that the electron energy varies as  $1/m_e^*d^2$ .



**Fig. 18.6.** Band structure of bulk silicon calculated by a highly accurate *ab initio* method (the GW approximation [12]) and also the tight-binding method [13]. Eigenenergies are plotted as a function of the wave vector  $\mathbf{k}$  in several directions of the Brillouin zone. The top of the valence band  $\varepsilon_v$  is taken as the zero energy. The width  $\varepsilon_c - \varepsilon_v$  of the band gap is 1.1 eV

is too simplistic and because boundary conditions describing real surfaces cannot be applied.

To establish the limits of this method, it is useful to consider the band structure of a bulk semiconductor, e.g., silicon, as illustrated in Fig. 18.6. The figure shows the electron energy as a function of the wave vector  $\mathbf{k}$ , also known as the dispersion relations. In the effective mass approximation, these dispersion relations are easily obtained by writing the kinetic energy as a function of  $\mathbf{k}$  and using the de Broglie relation  $\mathbf{p} = \hbar\mathbf{k}$ . The energy thus varies quadratically as a function of  $k = |\mathbf{k}|$ , as  $\varepsilon_c + \hbar^2 k^2 / 2m_e^*$  for the electrons and as  $\varepsilon_v - \hbar^2 k^2 / 2m_h^*$  for the holes. (The minus sign arises because a hole corresponds to an electron removed from an electron level.) From Fig. 18.6 it is clear that this quadratic approximation is only valid in the immediate vicinity of the band edges ( $\varepsilon_c$  and  $\varepsilon_v$ ) and that it generally overestimates the dispersion of the bands as one moves away from the band edges. As a consequence, the effective mass approximation overestimates the confinement energy. In the strong confinement regime, other more accurate methods are therefore used, such as empirically determined pseudopotentials and tight-binding, which describe the band structure throughout the Brillouin zone and over a wide range of energies ( $\approx 5$ – $15$  eV) around the band gap. Moreover, these techniques explicitly account for the atoms in the system, which is not the case in the effective mass approximation. A recent review of these methods can be found in [15]. We shall now outline what is involved.

### Empirical Pseudopotentials

This method is based on an expansion of the electron eigenfunctions in a plane wave basis, viz.,

$$\phi_i^{\text{e,h}} = \sum_{\mathbf{k}} c_i^{\text{e,h}}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{r}} , \quad (18.13)$$

where  $c_i^{\text{e,h}}(\mathbf{k})$  are treated as variational parameters. In practice, the basis of plane waves required to describe the lowest energy electron states is cut off at a maximal value of  $|\mathbf{k}|$  which fixes the minimal wavelength  $\lambda$  of oscillations in the wave function ( $|\mathbf{k}| = 2\pi/\lambda$ ). One difficulty arises because the basis in terms of which the wave functions are expanded must produce both the core and the valence states of the atoms. In reality, the core states in a solid or a molecule are very close to those of free atoms. The theory of pseudopotentials generates ways of eliminating core states from the calculations so that one may concentrate on the valence states, which are easier to write down. Consider a Schrödinger equation of the form

$$\left( -\frac{\hbar^2}{2m} + V \right) |\phi\rangle = \varepsilon |\phi\rangle . \quad (18.14)$$

As the eigenstate  $|\phi\rangle$  must be orthogonal to the core states  $|c\rangle$  produced by the same potential,  $|\phi\rangle$  is necessarily a highly oscillatory function in the vicinity of the atomic core, and this makes a description in terms of a plane wave basis impossible to achieve in practice. To get around this problem, it is useful to replace the potential  $V$  by a pseudopotential [16]:

$$V_{\text{ps}} = V + \sum_c (\varepsilon - \varepsilon_c) |c\rangle \langle c| . \quad (18.15)$$

The Schrödinger equation becomes

$$(T + V_{\text{ps}}) |\psi\rangle = \varepsilon |\psi\rangle . \quad (18.16)$$

The eigenvalues of this equation are equal to the energies of the valence states  $|\phi\rangle$ .  $V_{\text{ps}}$  is a complex operator which has no unique definition, since one may add any linear combination of core orbitals without changing the eigenvalues. This property allows one to optimise the pseudopotential in such a way that the pseudofunctions  $|\psi\rangle$  are as monotonic as possible. The pseudopotentials can be obtained by ab initio calculations and they can be optimised for application to a wide range of different systems (transferability criterion).

In the empirical pseudopotential approach, it is assumed that  $V_{\text{ps}}$  can be written as a sum of atomic contributions

$$V_{\text{ps}}(\mathbf{r}) = \sum_j v_j(\mathbf{r} - \mathbf{R}_j) . \quad (18.17)$$

The matrix of the potential in the plane wave basis is then written as a function of a small number of parameters which are subsequently fitted to obtain the best possible description of the electron structure of the system under investigation, usually the band structure of a crystalline solid. The valence and

conduction states of covalent semiconductors are generally well described by this method with atomic pseudopotentials that are relatively simple to implement. These pseudopotentials can then be transferred to handle other systems. Examples of applications of this method to semiconductor nanostructures and heterostructures can be found in [17].

### Tight-Binding Method

The tight-binding method describes the wave functions of the molecule as a linear combination of atomic orbitals [10, 11, 13]:

$$\phi = \sum_{j,\alpha} c_{j,\alpha} \varphi_{j,\alpha} , \quad (18.18)$$

where  $\varphi_{j,\alpha}$  is the orbital  $\alpha$  of the atom  $j$  at position  $\mathbf{R}_j$ . In general, a minimal basis of atomic orbitals is used, where minimality means that it is limited to the valence orbitals of the atom. For example, for silicon, one uses the  $3s$ ,  $3p_x$ ,  $3p_y$ ,  $3p_z$  orbitals ( $sp^3$  basis), and for hydrogen, the  $1s$  orbital. The tight-binding approximation consists in neglecting overlaps between orbitals, i.e., one assumes that the atomic orbitals are all mutually orthogonal. The allowed energies of the system are then given by diagonalising the matrix  $H$  of the Hamiltonian  $\hat{H}$  giving the eigenvalues

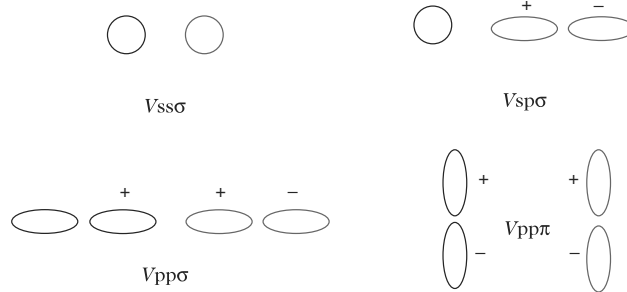
$$\det|H - \varepsilon I| = 0 , \quad (18.19)$$

where  $I$  is the identity matrix. The matrix  $H$  describing the Hamiltonian contains two types of term:

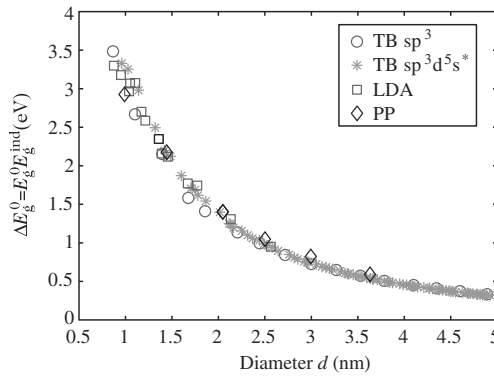
- Diagonal or intra-atomic terms  $H_{i\alpha,i\alpha} = \langle \varphi_{i\alpha} | \hat{H} | \varphi_{i\alpha} \rangle$ , which describe the energy of the orbital labelled  $\alpha$  of the  $i$ th atom in the system.
- Interatomic terms  $H_{i\alpha,j\beta}$  corresponding to the case where the two atoms labelled  $i$  and  $j$  are different. In general, only the terms between first, second, and possibly third nearest neighbours are included. The remaining elements are often simplified using the two-centre approximation which consists in neglecting the effect of the potential of any atom other than  $i$  and  $j$  on the matrix element  $\langle \varphi_{i\alpha} | \hat{H} | \varphi_{j\beta} \rangle$ , as in the case of a diatomic molecule. By making use of the symmetries of the problem, these interatomic elements can be expressed in terms of a limited number of independent parameters. For example, for atoms described in a basis ( $s$ ,  $p$ ), the various interactions now comprise only four terms (see Fig. 18.7):

$$V_{ss\sigma} , \quad V_{sp\sigma} , \quad V_{pp\sigma} , \quad V_{pp\pi} , \quad (18.20)$$

where  $\sigma$  (resp.  $\pi$ ) denotes a  $p$  orbital parallel to (resp. perpendicular to) the bond.



**Fig. 18.7.** Interactions between  $s$  and  $p$  orbitals in the tight-binding approximation



**Fig. 18.8.** Confinement energy in silicon nanocrystals as a function of their diameters, obtained by different ab initio and semi-empirical techniques: TB = tight-binding, LDA = ab initio method, PP = empirical pseudopotentials

Simple rules for the nearest-neighbour interaction potentials were obtained by Harrison [18]. The valence band of semiconductors is fairly well reproduced with these parameters. However, the conduction band is not. Approaches used to remedy this problem consist either in increasing the range of the interactions to second or even third nearest neighbours, or in increasing the size of the basis by adding a second  $s$  orbital, called the  $s^*$  orbital, or  $d$  orbitals. A very good description of the whole band structure can then be obtained, as shown in Fig. 18.6 for silicon.

Figure 18.8 shows the results obtained by this method for spherical silicon nanocrystals passivated by hydrogen. The calculation used tight-binding parameters from [13] in an  $sp^3$  basis. The confinement energy increases continuously as the size decreases. It behaves as  $1/d^2$  in the large nanocrystal limit (weak confinement), as predicted by the effective mass approximation.

It is interesting to compare these results obtained by a tight-binding model using an  $sp^3$  basis with other semi-empirical and ab initio methods. Figure 18.8 graphs the confinement energy as a function of the diameter  $d$  of

the nanocrystals for the  $sp^3$  model and an  $sp^3d^5s^*$  model [19]. The results obtained from a calculation based on semi-empirical pseudopotentials [20] and an ab initio technique [21] in the local density approximation (see below) are also graphed. Very good agreement is obtained between the results of the  $sp^3$  tight-binding model and the semi-empirical pseudopotential model in the range  $1 < d < 4$  nm, and also with the local density results for small nanocrystals. Finally, the agreement with the other tight-binding model, using an  $sp^3d^5s^*$  basis, is excellent over the whole size range. Many recent studies show that the semi-empirical methods give equivalent results (and equivalent to the ab initio methods), provided that they propose comparable descriptions for the bulk solid material and boundary conditions. On the other hand, the effective mass approximation overestimates the confinement energies by about 25% for  $d < 8.5$  nm and 50% for  $d < 4.5$  nm [13].

Semi-empirical methods such as pseudopotentials and tight-binding can be used to calculate many physical properties: optical properties, including excitonic effects [22] or electron–phonon coupling [23], dielectric properties [24] or transport properties such as the electrical spectroscopy of nanocrystals by scanning tunneling microscopy [25].

Another important field of applications for the semi-empirical methods is magnetism on the nanoscale, in particular in transition metal nanostructures [26], such as multilayer systems or clusters. In these systems, magnetic properties are very sensitive to the geometry and chemical environment of the atoms. Simulation then provides a way of understanding the complex phenomena involved. In the spirit of the tight-binding method, the spin–orbit interactions  $\xi_{ij}\mathbf{s}_i\cdot\mathbf{l}_j$  and spin–spin interactions  $J_{ij}\mathbf{s}_i\cdot\mathbf{s}_j$  between spins  $\mathbf{s}_i$  and/or orbital angular momenta  $\mathbf{l}_j$  associated with electrons are described by empirical parameters  $\xi_{ij}$  and  $J_{ij}$  which must be fitted in a suitable manner. It is important to note that the values of  $\mathbf{s}_i$  and  $\mathbf{l}_j$  depend heavily on the environment and cannot generally be taken as those in the bulk material. The Hamiltonian thus depends explicitly on the electronic state of the system, which itself depends on the Hamiltonian. One must then use the so-called self-consistent techniques (see below) which make the calculations rather more complex.

### 18.2.3 Ab Initio Methods

The empirical electronic approaches described above are extremely useful in the sense that they can be used to study systems containing large numbers of atoms. However, they have the disadvantage of being based on parameters that must be fitted for each system under investigation. This fitting requires a priori knowledge of the system, something that is not always available. Suppose for example that we wish to study the electronic properties of small silicon clusters. We can try to use the parameters fitted for bulk silicon, but can we be sure that these parameters are well suited to the Si–Si bond in



small clusters where the crystal environment is very different from the one encountered in bulk silicon?

The same question applies to the subject of the empirical potentials described in the first section of this chapter. Can an interatomic potential fitted to interactions between carbon atoms in diamond be used (transferred) to describe interatomic interactions in graphite, fullerenes, or carbon nanotubes? This problem of transferability of parameters from one system to another is a major obstacle which limits the predictive power and reliability of the empirical approaches.

To solve this problem, other approaches, known as *ab initio* approaches insofar as empirical parameters are no longer a prerequisite, have been developed. We shall now outline one widely used approach going by the name of density functional theory.

### Density Functional Theory (DFT)

The fundamental result from this theory is a formal demonstration that the energy of a system of ions and electrons can be written in the form of a functional of the electron charge density  $n(\mathbf{r})$  of the system, i.e.,  $E_0 = E_0[n(\mathbf{r})]$ . It is important to note that, in the general case, the energy is a functional of the  $N$ -body wave function of the system, i.e.,  $E_0 = E_0[\psi] = \langle \psi | \hat{H} | \psi \rangle$ . The density functional theory (DFT) thus provides a way of going from a total energy which depends explicitly on the  $3N_e$  variables describing the positions of the  $N_e$  electrons in the system to a functional depending only on the scalar field  $n(\mathbf{r})$ .

DFT thus shows that it is in principle possible to formulate an exact mean field approach. Indeed, the charge density  $n(\mathbf{r})$  is calculated as an average over the positions of all the electrons. One thus avoids the  $N$ -body problem which requires explicit knowledge of the wave function  $\psi(\mathbf{r}_1, \dots, \mathbf{r}_{N_e})$  and the resulting equations in  $3N_e$  dimensions. This result, which formalises an approximation previously known as the Thomas–Fermi approximation [27], earned one of its inventors, W. Kohn, the Nobel Prize for Chemistry in 1998 [28].

### Hohenberg–Kohn Theorem

Here we shall give the original formulation by Hohenberg and Kohn [29], presented in their seminal paper in 1964. In the notation of that article, the ionic potential acting on the electrons is called the external potential  $V^{\text{ext}}$ , in the sense that the ions are considered to be external to the electron system.

**Theorem 1.** (HK) *The potential  $V^{\text{ext}}$  is determined up to an additive constant by the electron density  $n(\mathbf{r})$ .*

It is clear that the ground state of an electron system is completely determined by the external potential  $V^{\text{ext}}$  and the number of electrons  $N_e$ . Given  $V^{\text{ext}}$  and  $N_e$ , the Hamiltonian  $\hat{H}$  is known and hence so are  $\psi$  and  $n(\mathbf{r})$ . What the first HK theorem shows is that the converse is also true, i.e., knowledge of  $n(\mathbf{r})$  determines the external potential  $V^{\text{ext}}$  uniquely, up to an additive constant.

Proof proceeds by reductio ad absurdum. Suppose there are two external potentials  $V_1^{\text{ext}}$  and  $V_2^{\text{ext}}$  which generate the same charge density  $n(\mathbf{r})$  everywhere in space, but two different wave functions  $\psi_1^{\text{GS}}$  and  $\psi_2^{\text{GS}}$  for the ground state. The variational principle applied to the Hamiltonian  $\hat{H}_1$ , viz.,

$$\hat{H}_1 = T + V^{\text{ee}} + V_1^{\text{ext}}, \quad (18.21)$$

where  $V^{\text{ee}}$  is the interelectron potential, leads to the inequality

$$E_1^{\text{GS}} < \langle \psi_2^{\text{GS}} | \hat{H}_1 | \psi_2^{\text{GS}} \rangle = \langle \psi_2^{\text{GS}} | \hat{H}_2 | \psi_2^{\text{GS}} \rangle + \langle \psi_2^{\text{GS}} | V_1^{\text{ext}} - V_2^{\text{ext}} | \psi_2^{\text{GS}} \rangle,$$

and hence,

$$E_1^{\text{GS}} < E_2^{\text{GS}} + \int d\mathbf{r} n(\mathbf{r})(V_1^{\text{ext}} - V_2^{\text{ext}})(\mathbf{r}). \quad (18.22)$$

We are considering here systems with non-degenerate ground state, whence the inequality is strict. Likewise, swapping the subscripts 1 and 2,

$$E_2^{\text{GS}} < E_1^{\text{GS}} + \int d\mathbf{r} n(\mathbf{r})(V_2^{\text{ext}} - V_1^{\text{ext}})(\mathbf{r}). \quad (18.23)$$

Adding the two inequalities,

$$E_1^{\text{GS}} + E_2^{\text{GS}} < E_2^{\text{GS}} + E_1^{\text{GS}}, \quad (18.24)$$

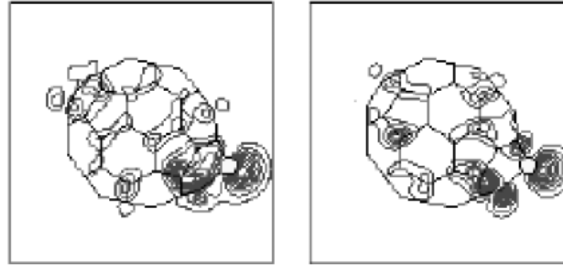
which is absurd. Hence, if  $V_1^{\text{ext}}$  and  $V_2^{\text{ext}}$  generate the same charge density  $n(\mathbf{r})$ , they must be equal.

This one-to-one relationship between charge density and external potential, which completely defines the system via the Hamiltonian  $\hat{H}$ , shows that the ground state energy of the system can indeed be written as a functional of the density.

### Kohn–Sham Equations

Furthermore, it can be shown that knowledge of the total energy as a function of the charge density allows one to write down a one-body Schrödinger equation known as the Kohn–Sham equation:

$$-\frac{\hbar^2}{2m_e} \nabla^2 \phi_n(\mathbf{r}) + V^{\text{ion}}(\mathbf{r}) \phi_n(\mathbf{r}) + V^{\text{eff}}[n](\mathbf{r}) \phi_n(\mathbf{r}) = \varepsilon_n \phi_n(\mathbf{r}), \quad (18.25)$$



**Fig. 18.9.** Electron distributions  $\phi^{\text{HOMO}}(\mathbf{r})$  (left) and  $\phi^{\text{LUMO}}(\mathbf{r})$  (right) associated with the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), respectively, of doped fullerene  $\text{C}_{59}\text{Si}$ . These states were obtained by solving the Kohn–Sham equations for the system

where  $V^{\text{ion}}$  describes ion–electron interactions, and the potential  $V^{\text{eff}}[n](\mathbf{r})$  is the mean potential which accounts for the action of the  $N - 1$  other electrons on a given electron. Like the total energy,  $V^{\text{eff}}[n](\mathbf{r})$  is a functional of the mean charge density of the system. It does not therefore depend explicitly on the positions  $(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_{N_e})$  of the electrons. The Kohn–Sham equation is a one-body equation that can be solved for systems comprising up to several thousand electrons. This equation is very widely used today. The eigenvalues  $\varepsilon_n$  and eigenfunctions  $\phi_n(\mathbf{r})$  form the basis for analysis of the electronic properties of solids and molecules in the DFT framework (see Fig. 18.9).

### Local Density Approximation (LDA)

The theorems dealing with the possibility of expressing the total energy  $E[n]$  and an effective interelectronic potential  $V^{\text{eff}}[n]$  as a function of the charge density  $n(\mathbf{r})$  are very formal. Exact analytic relations between these quantities are not known. In practice, one must therefore resort to approximations.

One widely used approximation is known as the local density approximation (LDA). This exploits the possibility of numerical solution of the exact  $N$ -body problem in the very specific case of a homogeneous system of interacting electrons, i.e., an electron gas with uniform charge density  $n$  in space. Indeed, the symmetry properties of this model system lead to a considerable simplification when solving the  $N$ -body Schrödinger equation. The energy  $E_{\text{hom}}(n)$  and potential  $V_{\text{hom}}^{\text{eff}}(n)$  can then be found very precisely for various densities  $n$  (‘hom’ stands for ‘homogeneous’).

The LDA uses these results, which are exact in the limit of a uniform charge distribution, to treat realistic systems in which the charge density  $n(\mathbf{r})$  varies in space. The idea is to say that the effective potential at a point  $\mathbf{r}$  of space with density  $n(\mathbf{r})$  is that of an electron gas with homogeneous density  $n = n(\mathbf{r})$ , so that  $V^{\text{eff}}[n](\mathbf{r}) = V_{\text{hom}}^{\text{eff}}(n(\mathbf{r}))$ . The hypothesis whereby

the effective potential at a point  $\mathbf{r}$  depends only on the charge density at that point (local character of the electron interaction) is an approximation which may seem rather crude, especially in covalent systems where the charge density can actually be extremely inhomogeneous. However, the LDA leads to good results: interatomic distance, binding energies, vibrational frequencies, etc., can be calculated to within an error of only a few percent compared with experiment, and this without any recourse to adjustable empirical parameters. Other approximations called gradient correction approximations bring in not only the charge density, but also its gradient, in order to account more accurately for spatial charge variations.

### Self-Consistency

A crucial feature in ab initio calculations is the idea of self-consistency. In order to find the wave functions  $\phi_n$ , one must solve the Kohn–Sham equations with Hamiltonian  $H^{\text{KS}}$ . However, this Hamiltonian depends on the charge density, which is itself constructed from the wave functions  $\phi_n$  by

$$n(\mathbf{r}) = \sum_{n=\text{occ}} |\phi_n(\mathbf{r})|^2,$$

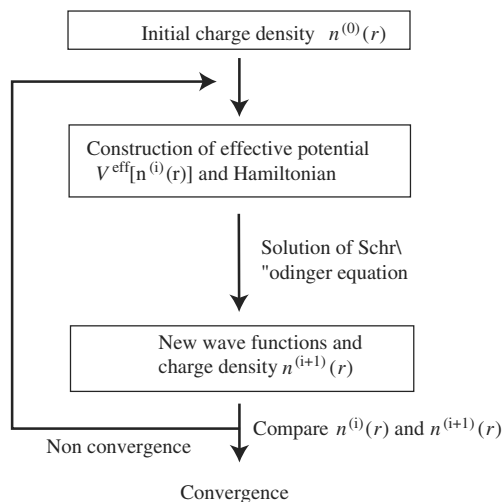
where the condition  $n = \text{occ}$  indicates that the sum is taken over occupied states. To solve this problem, an iterative self-consistent approach is implemented. Starting with an ‘arbitrary’ initial density  $n^0(\mathbf{r})$  (generally chosen as the superposition of the atomic charge densities), an initial Kohn–Sham Hamiltonian  $H^{\text{KS}}[n^0(\mathbf{r})]$  is constructed. This is inverted to give states  $\phi_n$  which can then be used to construct a new charge density  $n(\mathbf{r})$ . If  $n(\mathbf{r})$  is very different from  $n^0(\mathbf{r})$ , the Hamiltonian  $H^{\text{KS}}[n(\mathbf{r})]$  is constructed and solved to yield another charge density, and so forth. This algorithm is shown schematically in Fig. 18.10.

This self-consistent process is very important in particular in multi-element systems where there may be charge transfer from one chemical species to another (as in  $\text{NaCl} = \text{Na}^+\text{Cl}^-$ , for example). The self-consistent loop provides a way of following such charge transfer. These processes are also at work in the response of a system to an external perturbation. Let us consider the important case of the interaction between a cluster and light. A priori, first order perturbation theory [30] easily gives the variation  $\delta\phi_n$  of the orbitals as a function of the external perturbation  $\delta V^{\text{ext}}$  (in this example, an electromagnetic wave):

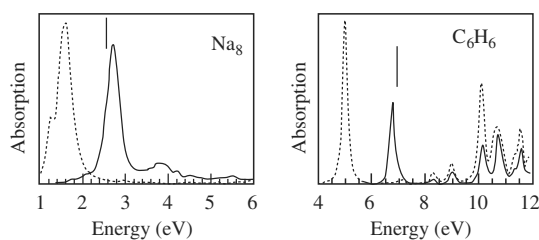
$$\delta\phi_n(\mathbf{r}) = \sum_{m \neq n} \frac{|\langle \phi_n | \delta V^{\text{ext}} | \phi_m \rangle|^2}{\varepsilon_n - \varepsilon_m} \phi_m(\mathbf{r}). \quad (18.26)$$

The associated variation in the charge density, viz.,

$$\delta n(\mathbf{r}) = \sum_{n=\text{occ}} \left[ \delta\phi_n^*(\mathbf{r})\phi_n(\mathbf{r}) + \phi_n^*(\mathbf{r})\delta\phi_n(\mathbf{r}) \right],$$



**Fig. 18.10.** Flow diagram for a self-consistent convergence loop



**Fig. 18.11.** Optical absorption spectra of  $\text{Na}_8$  (left) and benzene  $\text{C}_6\text{H}_6$  (right). Spectra with and without self-consistent effects are indicated by *continuous* and *dotted curves*, respectively. Experimental results for the main absorption peaks (the so-called Mie frequency for  $\text{Na}_8$  and the transition  $\pi \rightarrow \pi^*$  for benzene) are indicated by *vertical bars*. For time-dependent Hamiltonians, the density functional theory is called TD-DFT [6]

would then yield the polarisability of the cluster, for example. This approach is in fact incorrect. Indeed, when the charge density changes, the effective potential  $V^{\text{eff}}[n(\mathbf{r})]$  also changes. The perturbation is not the external field  $\delta V^{\text{ext}}$  alone, but the sum  $\delta V^{\text{ext}} + \delta V^{\text{eff}}[n]$ , which depends self-consistently on the charge variation. The difference between the two approaches is illustrated in Fig. 18.11 for the optical absorption spectra of a small metallic cluster ( $\text{Na}_8$ ) and the benzene molecule  $\text{C}_6\text{H}_6$ .

#### 18.2.4 Ab Initio Calculation of Interatomic Forces

Ab initio calculations give us the energy of a system, taking into account electron–electron energies (via the effective potential), ion–electron energies

and ion–ion energies. As we saw in the first section, energy calculations can be used to classify the most stable structures and work out the forces exerted on the ions in order to relax the atomic positions or carry out molecular dynamics.

Forces are calculated from the gradient of the total energy with respect to atomic positions, viz.,  $\mathbf{F}_I = -\nabla_{R_I} E$ . With empirical potentials, the explicit dependence of the energy on the positions  $\mathbf{R}_I$  made this derivation particularly straightforward. In the quantum case, only the ion–ion interaction energy depends explicitly on the position of the nuclei. In contrast, electron–electron and ion–electron energies depend only implicitly on ion positions through the charge density which is adapted self-consistently to the positions of the nuclei. Quite generally in quantum mechanics, the gradient of the total energy, viz.,

$$\nabla_{R_I} E = \nabla_{R_I} \langle \phi | \hat{H} | \phi \rangle = \langle \nabla_{R_I} \phi | \hat{H} | \phi \rangle + \langle \phi | \nabla_{R_I} \hat{H} | \phi \rangle + \langle \phi | \hat{H} | \nabla_{R_I} \phi \rangle, \quad (18.27)$$

depends on the gradient of the wave functions, which do not depend explicitly on the positions of the atoms. To get around this difficulty, forces are generally calculated using the so-called Born–Oppenheimer approximation. The typical time scale for electron dynamics is the femtosecond ( $10^{-15}$  s). Due to the small mass ratio of electrons to nuclei, the latter move between a hundred and a thousand times more slowly. This observation justifies the assumption that, when the nuclei move, the electrons are able to relax almost instantaneously (on the scale of ion dynamics) into the ground state corresponding to the instantaneous position of the atoms. As far as this approximation is valid, the electron wave function  $|\phi\rangle$  is therefore an eigenvector of the Hamiltonian  $\hat{H}$  with eigenvalue equal to the energy  $E_0$  of the ground state. Hence,

$$\langle \nabla_{R_I} \phi | \hat{H} | \phi \rangle + \langle \phi | \hat{H} | \nabla_{R_I} \phi \rangle = \langle \nabla_{R_I} \phi | \phi \rangle E_0 + E_0 \langle \phi | \nabla_{R_I} \phi \rangle = E_0 \nabla_{R_I} \langle \phi | \phi \rangle, \quad (18.28)$$

which is zero since, by normalization,  $\langle \phi | \phi \rangle$  is a constant equal to unity. This result, known as the Hellmann–Feynman theorem, is a special case of the variational principle. In practice, the fact that the variation of the wave functions drops out of the force expression is an important result. Note that this result does not apply when the electrons are excited out of their ground state by some (rapidly) time-varying external perturbation, such as an electromagnetic wave.

Knowing the forces derived from ab initio calculations (and hence without appealing to empirical potentials), one can carry out molecular dynamics for systems in which interactions are difficult to describe by a simple formula depending only on the relative positions of the atoms. This happens in particular for multi-element systems, where charge transfer from one atom to another can change during the dynamics, significantly altering the more or less ionic nature of the interactions.

### 18.2.5 Using Electron Wave Functions and Eigenvalues

Whether calculated semi-empirically or by ab initio methods, the wave functions and energies  $(\phi_i, \varepsilon_i)$  obtained above can be used to find many properties of the material. We have already mentioned the interaction with an external electric field. The variation of the wave functions [see (18.26)] can be used to obtain the time variation of the charge density, viz.,  $\delta n(\mathbf{r}, t)$ , and hence in particular, the dynamic polarisability  $\alpha_{ij}(\omega)$  for clusters, or the dielectric constant and plasmon modes for solids. These quantities are fundamental for studying the optical absorption by nano-objects or photonic nanocrystals. Likewise, by calculating the variation of the wave functions and electron energies in an applied magnetic field, one can infer the magnetic response properties of the material (see [2], Chaps. VIII–X).

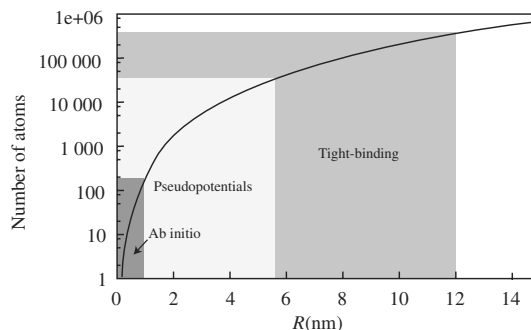
Another important example in the nanotechnology context is the possibility of modelling STM images. As discussed in Sect. 3.2.2 of Chap. 3, in the simple Tersoff–Hamann approximation, the tunneling current can be found to a first approximation from the electron charge density  $n(\mathbf{r})$  associated with the surface or the molecule interacting with the STM tip. More sophisticated approaches are also based on knowing the wave functions of the surface to be imaged and the STM tip just above it. The example of tunnel imaging is a special case of a much wider field, namely electron transport in nanostructures, a subject that can be tackled using modern numerical simulation techniques [13].

## 18.3 Conclusion

The examples discussed here should be sufficient to demonstrate the usefulness of numerical simulations for studying the geometry of nanostructures, as well as their electronic and optical properties. Many other physical quantities are commonly calculated by computer, e.g., superconducting transition temperature, magnetic anisotropy and moment, NMR or Raman spectra, etc., with accuracies to within a few percent.

Many approaches have been omitted here, due to lack of space. One could mention the kinetic or statistical Monte Carlo methods for structural properties, phase diagrams, or dynamics, and the quantum ab initio approaches such as the Hartree–Fock method, quantum Monte Carlo, or configuration interaction, used to study electron properties. The reader is referred to textbooks dealing specifically with these techniques [1].

The choice of one method or another is very often based on a compromise between reliability, accuracy, and computation time (see Fig. 18.12). Ab initio calculations are generally more reliable and accurate than the empirical approaches, but much more costly in terms of computer time. As an example, in order to follow a hundred atoms for just 10 ps using ab initio molecular dynamics, one would have to wait more than a month even with access to a



**Fig. 18.12.** Number of atoms in a spherical silicon nanocrystal as a function of its radius  $R$ . Current limits of the main techniques for calculating electron structure are indicated. Nanostructures commonly studied experimentally lie in the size range 2–15 nm

good work station. Over the same period, simulations using empirical potentials can handle several thousand atoms as they evolve over several hundred picoseconds.

Hence, despite the considerable recent development of ab initio methods, the semi-empirical techniques continue to play an important role in the analysis of structural, dynamic and electronic properties. They can be used to tackle problems where ab initio techniques remain too complicated to implement. They can also be used to obtain simple and pedagogical descriptions of complex problems. The fact remains, however, that the problem of choosing adjustable parameters often requires a quality control on the results obtained, by carrying out empirical and ab initio calculations in parallel on the same test system.

## References

1. Thijssen, J.M.: *Computational Physics*, Cambridge University Press (1999); Allen, M.P., and Tildesley, D.J.: *Computer Simulation of Liquids*, Oxford Science and Publications (1987)
2. Joyes, P.: *Les agrégats inorganiques élémentaires*, Les éditions de physique, Les Ulis (1990)
3. Kittel, C.: *Introduction to Solid State Physics*, 7th edn., Wiley, New York (1996) Chap. 7
4. Cohen-Tannoudji, C., Diu, B., Laloë, F.: *Quantum Mechanics*, Vol. II, Wiley, New York (1977)
5. Blase, X., and Jensen, P.: *Les Matériaux virtuels*, La Recherche **352**, 40–44 (April 2002)
6. Blase, X., and Ordejón, P.: Phys. Rev. B **69**, 085111 (2004)
7. Kittel, C.: *Introduction to Solid State Physics*, 7th edn., Wiley, New York (1996) Chap. 8



8. Bastard, G.: *Wave Mechanics Applied to Semiconductor Heterostructures*, Les éditions de physique, Les Ulis (1988)
9. Chelikowsky, J.R., and Cohen, M.L.: Phys. Rev. B **14**, 556 (1976); Cohen, M.L., and Chelikowsky, J.R.: *Electronic Structure and Optical Properties of Semiconductors*, Springer Series in Solid State Physics, Springer-Verlag, Berlin (1988); Wang, L.-W., and Zunger, A.: Phys. Rev. B **51**, 17398 (1995)
10. Kittel, C.: *Introduction to Solid State Physics*, 7th edn., Wiley, New York (1996) Chap. 9
11. Ashcroft, N., and Mermin, N.D.: *Solid State Physics*, Saunders College Publishing, International Edition (1976) Chap. 10
12. Reining, L.: private communication
13. Delerue, C., and Lannoo, M.: *Nanostructures. Theory and Modelling*, Springer-Verlag, Berlin, Heidelberg, New York (2004)
14. Grundmann, M., Stier, O., and Bimberg, D.: Phys. Rev. B **52**, 11969 (1995); Jiang, H., and Singh, J.: Phys. Rev. B **56**, 4696 (1997); Pryor, C.: Phys. Rev. B **57**, 7190 (1998)
15. Di Carlo, A.: Semicond. Sci. Technol. **18**, R1–R31 (2003)
16. Hamann, D., Schlüter, M., and Chiang, C.: Phys. Rev. Lett. **43**, 1494 (1979)
17. Wang, L.-W., Zunger, A.: J. Phys. Chem. **98**, 2158 (1994); Fu, H., Wang, L.-W., Zunger, A.: Phys. Rev. B **57**, 9971 (1998); Wood, D.M., Zunger, A.: Phys. Rev. B **53**, 7949 (1996)
18. Harrison, W.A.: *Electronic Structure and the Properties of Solids: The Physics of the Chemical Bond*, Dover Publications, New York (1989)
19. Jancu, J.M., Scholz, R., Beltram, F., and Bassani, F.: Phys. Rev. B **57**, 6493 (1998)
20. Zunger, A., and Wang, L.-W.: Appl. Surf. Sci. **102**, 350 (1996)
21. Delley, B., and Steigmeier, F.: Appl. Phys. Lett. **67**, 2370 (1995)
22. Franceschetti, A., Zunger, A.: Phys. Rev. B **62**, 2614 (2000); Martin, E., Delerue, C., Allan, G., Lannoo, M.: Phys. Rev. B **50**, 18258 (1994)
23. Allan, G., Delerue, C.: Phys. Rev. B **66**, 233303 (2002)
24. Wang, L.-W., Zunger, A.: Phys. Rev. Lett. **73**, 1039 (1994); Lannoo, M., Delerue, C., Allan, G.: Phys. Rev. Lett. **74**, 3415 (1995)
25. Bakkers, E., Hens, Z., Zunger, A., Franceschetti, A., Kouwenhoven, L., Gurevich, L., Vanmaekelbergh, D.: Nano Lett. **1**, 551 (2001); Niquet, Y.M., Delerue, C., Allan, G., Lannoo, M.: Phys. Rev. B **65**, 165334 (2002)
26. Pastor, G.M.: Theory of cluster magnetism. In: *Atomic Clusters and Nanoparticles*, ed. by C. Guet et al., Les Houches Session LXXIII EDP Sciences, Springer-Verlag, Berlin, Heidelberg (2001) p. 335
27. Kittel, C.: *Introduction to Solid State Physics*, 7th edn., Wiley, New York (1996) Chap. 10
28. Kohn, W.: Nobel lecture, Rev. Mod. Phys. **71**, 1255 (1999)
29. Hohenberg, P., and Kohn, W.: Phys. Rev. **136**, B864 (1964)
30. Cohen-Tannoudji, C., Diu, B., Laloë, F.: *Quantum Mechanics*, Vol. II, Wiley, New York (1977) Chap. 11

## Computer Architectures for Nanotechnology: Towards Nanocomputing

C. Gamrat

The advent of nanotechnology and improved control of fabrication processes have led to the elaboration of nanocomponents with novel and often surprising properties. Among all the fields of application that might benefit from this progress, there is one which seems to hold especially great promise: this is the area of data processing devices. The design of these future nanocomputers will come by proficient control of a particularly well-suited computer architecture. Indeed, this is an opportunity to raise a certain number of questions concerning the suitability of computer architectures in use today and to revisit the thinking behind the elements that make them up: memory cells, logic gates, and interconnects, but also information coding and advanced architectures. It is this whole new field of research on the frontier between components and systems that we aim to explore in the present chapter.

### 19.1 Introduction

Over forty years ago, when Bardeen, Brattain and Shockley invented the transistor, the applications imagined for this new device were very different to those we would imagine today for the same invention. At the time, the amplifying characteristics of the transistor and its low operating voltage naturally found applications in radio, low frequency amplification, measuring equipment, and other analog electronic devices. But today, whenever a new electronic device becomes available, one no longer thinks of the record player, the radio, or the television as a primary application, but rather of the computer. Most electrical signals representing sounds, pictures, telephone conversations, right down to the spelling mistakes I constantly make when writing this text, are digitised. For this reason, the computer, fast and skillful manipulator of binary data, is naturally placed at the centre of modern technology, for which it has become the driving force. This has become so true today that we can never be completely sure where a computer might be hiding, even in our own homes! It may be in our portable telephone, in the least expected corners of

our car, in our television set, the washing machine, or even the coffee percolator! The omnipresence of this device has become so important that we can no longer imagine an application of technology without calling for its assistance.

So at a time when nanotechnology promises new elementary components, it is quite natural to ask how these may be put to use to build a computer. As the previous chapters have illustrated, nanoscience offers a great diversity of fabrication techniques and operating principles which exploit the basic physical phenomena. Amongst this diversity, it is hard to identify those devices that might be useful in realising a computer, or part of one. Apart from considerations of function and performance with regard to the basic device, other aspects must be reviewed, such as questions of fabrication, compatibility with other technologies, and cost. Indeed, the last point tends to become a key feature in the strategic field of computer technology. Of course, given the wealth of subjects that could be subsumed under the theme of this chapter, there will be no attempt to be exhaustive. The reader should consider it as an introduction to the general problem situation, revealing the subject in a light that he or she may not previously have imagined.

In the following, we shall not go into details concerning the nanodevices that are likely to be used to build the future nanocomputer. Indeed, the physical characteristics of these devices together with a description of how they can be made have already been discussed at length in the various chapters of this book. We shall be concerned here with describing the basic functions and the way they can be put together in order to make a data processing system. We shall also try to identify those approaches that look most likely to contribute to the architecture of the future nanocomputer. The guiding idea here will be to provide the reader with a system-based overview of the problem, so that he/she may imagine the kind of complex system one might set out to build with a set of nanocomponents. We shall also attempt to answer certain questions relating to the main architectural principles underlying today's computers. Will these principles, so clearly validated by current microelectronic technology, maintain their relevance in the novel context provided by nanotechnology? Is not the advent of a new technology an opportunity to rethink at least some of these principles? Can we expect qualitative or quantitative advantages over the best solutions known today?

There is no doubt that the following discussion will raise more questions than it will answer. The fact is that this new field remains to be explored, and suitable architectural solutions remain to be imagined, bearing in mind that no one today has the blueprint for a working nanocomputer tucked away in their drawer!

Section 19.2 begins with a brief review of the current situation and trends in the digital technologies, before identifying the typical computer architecture and the indispensable basic elements that must be mastered in order to build it. We then discuss the critical points and solutions that could be brought to bear by nanotechnology in the familiar architectural framework.

In Sect. 19.3, we review several proposed architectures which seem well-suited to the specific context of nanotechnology. Some of these ideas have never received much attention against the background of conventional technology and we shall see how far nanoscience might be able to rehabilitate them.

Finally, in Sect. 19.4, we shall turn to several important points concerning system design, such as reliability and pooling of resources in heterogeneous computation.

## 19.2 Computer Architecture and Basic Functions

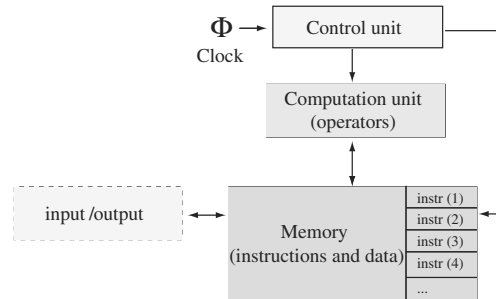
Between a technology capable of producing a range of elementary devices and its assembly in view of some specific application, one enters the province of architecture, whose final aim is to integrate these devices into a usable system. At the same time, the computer architect must also propose orientations for the development of technological components that are consistent with the needs of a system, and hence, at the end of the day, those of its user. At the crossroads between technology and its use, architecture, or the art of assembling the building blocks to make a useful ensemble, is therefore an inescapable crossing point.

### 19.2.1 Typical Architecture of a Computer

To identify the main functions, let us review certain features characterising the structure of the typical computer. In fact, the archetypal computer with which we are so familiar today is based upon a principle stated by A. Turing in 1935, when he first described the famous machine which now carries his name. This purely conceptual invention was developed at the time to illustrate his mathematical research in the field of logic, undecidability, and other concepts in computation theory. The idea was taken up and adapted ten years later by J. von Neumann and the result became the foundation stone of computer architecture as we know it. So what exactly is this idea?

In this architecture, a processing unit is connected to a memory device which stores the data to be processed and also a list of processing instructions. A control unit synchronises read/write operations between these two elements. To be complete, some device must handle communications with the outside world. Figure 19.1 illustrates this concept in a simple, schematic way.

1. Read an instruction ( $n$ ) in memory.
2. Read the necessary data if any.
3. Carry out the instruction.
4. Write the result in memory.
5. Move forward to the next instruction ( $n + 1$ ).
6. Go back to step 1 and continue until the instruction ‘stop’.



**Fig. 19.1.** Highly simplified view of how a computer works

This architecture is characterised on the one hand by the fact that processing is sequential, and on the other by the fact that data and program instructions are stored together in the same memory. This was the key refinement made to the Turing machine by von Neumann. Not all instructions that are executed actually process data. For example, the so-called control instructions serve to manipulate the sequence itself, effecting operations such as omission, branching, or stopping. These essential instructions are used to write more complex algorithms than a simple linear series of instructions.

Memories, interconnects and operators are thus the basic elements for building a computer according to the model we have just described. We may note in passing that, among these three functions, only one – the operator – can modify the state of the data. Memories and connections must in no way alter the state of the data they handle.

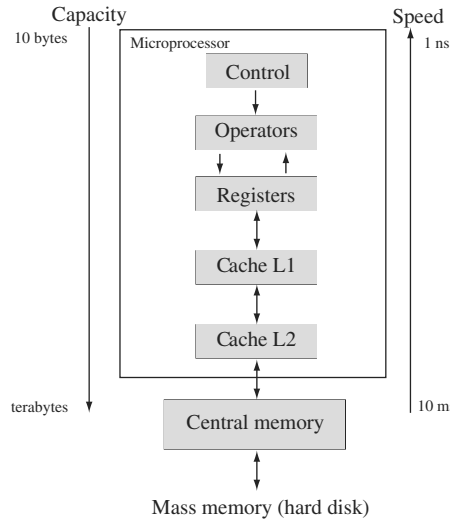
### 19.2.2 Memory

The role of memory is simply to conserve a particular state at a given instant of time and to be able to reproduce it faithfully when required. The essential characteristics of a good memory element are:

- access time (read and write),
- storage density per unit area,
- data retention time.

The way memory elements are used within a typical computer architecture must fit in with the specific properties of the available memory technologies. Indeed, between the core of the processor which actually processes data and the central memory in which the data is stored, the required data and instructions pass through a succession of layers each of which may have different density and speed characteristics. This hierarchical structure of the memory is illustrated in Fig. 19.2.

In the core of the processor, a sequence of instructions is executed at high speed over a set of registers. The latter are small scale memories, containing



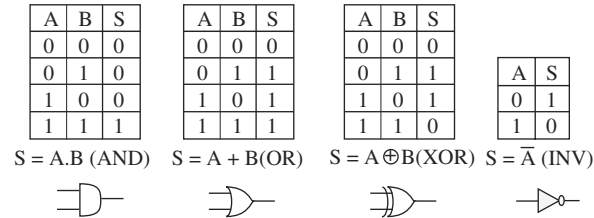
**Fig. 19.2.** Typical memory hierarchy

**Table 19.1.** Speeds and storage capacities at various levels in the memory hierarchy

Level	Technology	Speed	Capacity
Register	Same as core (volatile)	0.5 ns	10–1 000 byte
Cache L1	Same as core (volatile)	1–5 ns	1–100 Kbyte
Cache L2	Static RAM (volatile)	3–10 ns	1–10 Mbyte
Central memory	Dynamic RAM (volatile)	5–30 ns	1–10 Gbyte
Mass memory	Hard disk (non-volatile)	10 ms	1 Tbyte

several tens of bytes each and operating at the same speed as the processor core. To transfer the data from the register level to the central memory, intermediate memories known as cache memory are interposed. Their role is to adapt the speed of the processor core to that of the central memory. This hierarchy in the memory structure arises because different memory technologies are optimal either for the speed or for the storage density. This is shown in Table 19.1 for the memory hierarchy of a typical PC using an Intel Pentium processor.

Let us note in passing an opportunity for radically simplifying and improving computer architectures. Indeed, a technology able to achieve compact and economical memories with a capacity of several gigabytes and with an access speed of nanosecond order would make it possible to omit the cache memories.



**Fig. 19.3.** Basic logic operators (gates)

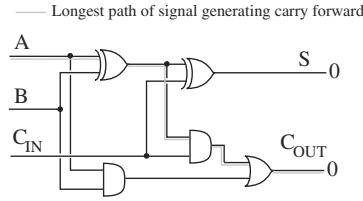
### 19.2.3 Interconnects

The role of interconnects (wires) is simply to transport a binary state from one place to another. Although this function may sound trivial, it is no less essential and critical. In the context of conventional electronic technology, connections are directly characterised by the electron transport properties of the materials used to make lines. Depending on the geometric parameters of the metals employed, connection lines are primarily characterised by properties such as resistivity ( $1.7$  and  $2.7 \times 10^{-8} \Omega \text{m}$  for Al and Cu, respectively), or the capacitance and inductance per unit length. Moreover, depending on the environment in which the lines are used (geometry, dielectric) and the type of signal they transport, mutual interaction parameters can become relevant, even critical. These properties have a direct impact on the static and dynamic behaviour of the lines. As a general rule, the more cramped the geometry, the more important it is to have high quality interconnects. Furthermore, given the ever-increasing operating frequencies, particularly in processor cores, it is important to consider signal propagation speeds. Indeed, this parameter was neglected in the past when designing integrated circuits, but with a clock frequency at 3 GHz, the distance travelled by a signal in a half-period is reduced to a few centimeters.

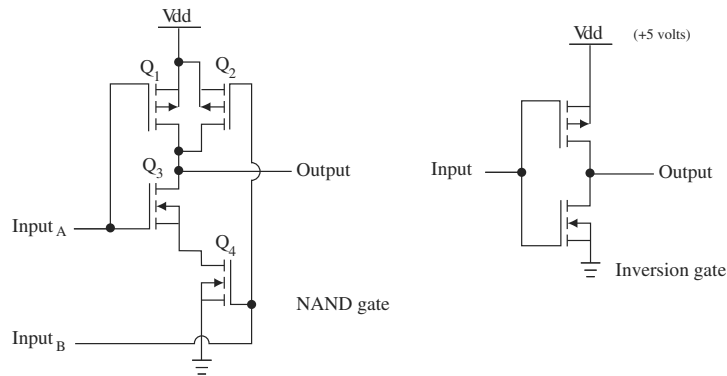
### 19.2.4 Operators

Computation operators constitute the active part of the computer. This is where useful data is manipulated and transformed. The physical implantation of computation operators is closely connected with the way one chooses to code information. Although binary coding is most familiar to us today, this has not always been the case. However, Boolean logic and handling of binary states are perfectly suited to the microelectronics technologies. The elementary operators for binary coding are shown in Fig. 19.3.

The many combinations of these basic operators can achieve more complex functions by application of Boolean algebra. For example, a full 2-bit adder can be built as shown in Fig. 19.4. This very simple example clearly brings out the notion of a circuit upon which the more complex operators are based. For example, generation of the carry signal involves passage through many elementary gates. A signal restitution mechanism is required. In microelectronics



**Fig. 19.4.** A 2-bit adder



**Fig. 19.5.** Physical implantation of operators (CMOS technology)

technology, it is the transistor, elementary component of the logic gate, which fulfills this role. Arbitrarily complex circuits can thus be constructed. With other technologies, it may grow difficult to make computer elements by physically combining circuits unless appropriate signal restitution or refreshment devices become available.

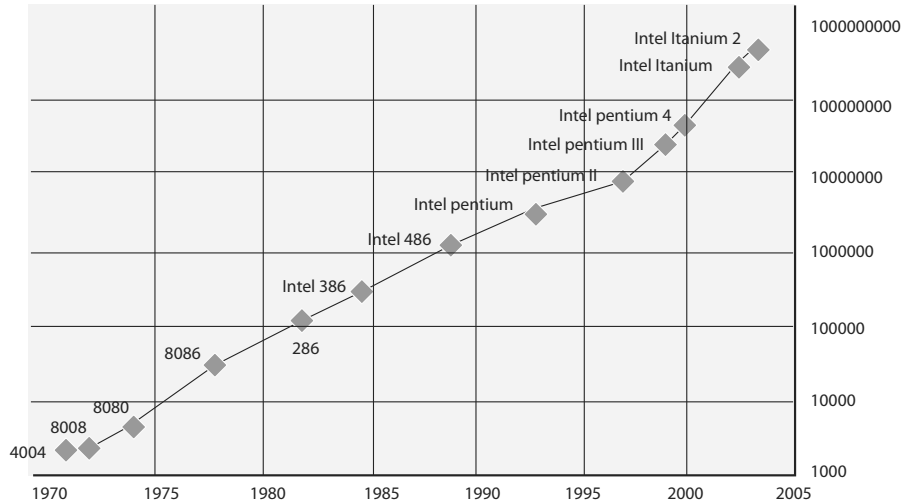
### 19.2.5 Technological Considerations

With the electronic technologies we know so well, the three functional elements just described can be produced in various ways. For example:

- A memory cell using a condenser which stores charges, or based on a so-called bistable circuit like an Eccles–Jordan flip-flop circuit.
- An interconnect line made from a simple metal conductor, allowing electron charges to circulate. This is the simplest element to implement, provided that speed and miniaturisation constraints are not too demanding.
- Operators are made up of simple switches. Suitably arranged and adjusted, they can reproduce all possible Boolean logic functions. The role of the switch is generally entrusted to a transistor. Figure 19.5 shows how NAND and NOT operators can be built, for example.

Ever since the advent of microelectronics at the beginning of the 1970s, these elements have been made in the form of integrated circuits. These more and





**Fig. 19.6.** Moore's law (according to Intel). The *vertical axis* shows the number of transistors per processor

more complex circuits bring together all the elements required for a computer on one chip: control and computation units, memory hierarchy and input/output devices. Considerable progress has been made in microelectronics technology. Moore's law [3], an empirical law due to G. Moore, one of the founding fathers of Intel, states that the complexity and performance of circuits based on microelectronics technology will double every 18 months (see Fig. 19.6).

This law has been so well observed since the 1970s that the microelectronics industries use it to determine their industrial and economic predictions. At the present time, what preoccupies manufacturers in this sector is precisely the question as to whether Moore's law will continue to be so exactly satisfied in the years to come, given that the well-being of their business depends on it. As an illustration of Moore's law and its impact on how the products have evolved, Table 19.2 gives an overview of the development of the Intel processor over a period of 30 years.

The reduction in characteristic transistor sizes and the descent to nanometric dimensions will inevitably raise a certain number of physical problems which are bound to reduce the slope of the Moore curve. Among such limitations, we have already mentioned the increasing contribution of interconnects. There is another difficulty with regard to the power density of circuits and the technological means available for evacuating the heat they generate. Yet another problem is to run devices in which the useful dynamics continues to approach the background noise level. These difficulties are already discussed in technological forecasts by the microelectronics industry [28], and some solutions have been envisaged to extend the lifetime of current technology (stressed

**Table 19.2.** Microprocessor characteristics over the last 30 years. Note that most of the transistors in the Pentium 4 are used in the cache memory zones

Name	Year	Number of transistors	Technology [ $\mu\text{m}$ ]	Clock [MHz]	Data bus [bits]	MIPS
8080	1974	6 000	6	2	8	0.64
8088	1979	29 000	3	4.77	16/8	0.5
80286	1982	134 000	1.5	6	16	1
80386	1985	275 000	1.5	16	32	5
80486	1989	1 200 000	1	33	32	25
Pentium	1993	3 100 000	0.8	75	32/64	100
Pentium II	1997	7 500 000	0.35	233	32/64	300
Pentium III	1999	9 500 000	0.25	450	32/64	510
Pentium 4	2000	42 000 000	0.18	1 500	32/64	1 700
P4 (Prescott)	2004	125 000 000	0.090	3 600	32/64	7 000

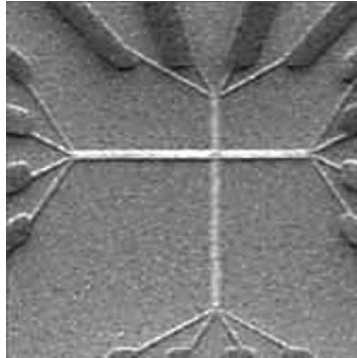
silicon, high permittivity dielectrics), whilst the search continues for new devices (SET, molecular electronics). In a word, the industry is already moving gradually into the age of nanoelectronics.

### 19.2.6 Nanomemories, Nano-operators, Nanoconnections

The study of nanotechnological devices presented in previous chapters of this book has revealed several possibilities for making the elements required to construct a hypothetical nanocomputer. Among these, carbon nanotube (CNT) technology looks particularly promising. Metallic or semiconducting CNTs seem well-suited to play the part of both interconnects and transistors. Indeed, the feasibility of FET-type transistors using semiconducting CNTs has already been demonstrated [34]. Combining CNTs into a crossbar-type structure, Lieber et al. in Harvard have suggested making memory cells that work via a simple phenomenon of electrical repulsion and attraction [24]. The point of such a structure would be to achieve (theoretical) switch densities of the order of  $1\,000\text{--}10\,000\ \mu\text{m}^{-2}$ , i.e.,  $1\text{--}10\ \text{Gbit}/\text{mm}^2$ .

However, even when it becomes possible to make such matrices, other problems will remain before they can be put to use. To begin with there is a very real problem of scale which is illustrated in Fig. 19.7. An array of 64 active points fabricated at the intersection of two bundles of eight nanowires (HP Labs) occupies an area of about  $1\ \mu\text{m}^2$  (central region of the figure). In stark contrast, the copper or aluminium conductors giving access to this array (periphery) occupy several hundred  $\mu\text{m}^2$ .

Hence, nanotechnology seems to supply all the building blocks required to build a computer as we know it, and yet the von Neumann architecture is perhaps not the best suited to the nanoscale world. First of all, the computer architecture presented at the beginning of this chapter is organised into



**Fig. 19.7.** Micro–nano interconnects. Courtesy of HP Labs

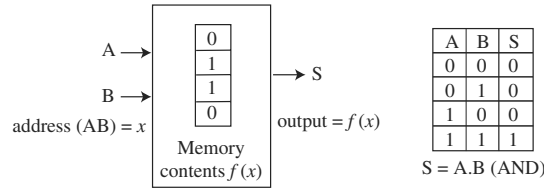
well-defined blocks. Such an organisation leads inevitably to lengthy non-local interconnects and implies a structure with a high level of granularity. In contrast to the memory element, the control unit and operator elements possess a rather irregular internal structure. This irregularity in the distribution of interconnect circuits complicates the timing control of signals within an essentially synchronous computer architecture. Finally, the computer made in this way is very vulnerable to defects. Whether one is dealing with fabrication defects, breakdown, or dynamical perturbation due to noise, this architecture is not robust in the face of mishap. Certain elements like the control unit are especially concerned here. It is therefore wise to seek an alternative architecture better suited to the nanoworld. This is what we shall explore in the next section.

## 19.3 Some Ideas for a New Architecture

### 19.3.1 Calculating with Memory Alone

A great deal of work has focused on the fabrication of memory cells. Indeed, it seems relatively natural to conceive of a bistable, even multistable device, which can be initialised in a particular state, then design a circuit for reading this state. Many physical phenomena exhibit stable states – static or dynamic equilibrium states – that could perhaps be exploited for this purpose. Hence, the memory function is likely to be the prime beneficiary of advances in nanotechnology.

Let us imagine a technology that could be used to obtain gigantic and very dense memory planes (several Gbytes/mm<sup>2</sup>), but with which it would be difficult to build logic element circuits of the kind described above. With this technology, would it be possible to carry out operations other than data storage?



**Fig. 19.8.** Look-up table (LUT), exemplified by the AND function

A simple answer goes by the name of the look-up table (LUT). This extremely simple device can be used to make any operator simply by initialising a memory zone with the states of its truth table. For example, consider a memory containing 4 binary cells that can be separately addressed by means of a 2-bit code via the AB signals (see Fig. 19.8). Now add a connector (S) to the memory in order to output the states contained within the addressed cells, and fill the cells of the memory with the states corresponding to the function  $f(AB)$  that one hopes to realise. By applying a value  $x$  to the address input of the memory, the value  $f(x)$  is output. In brief, the truth table of the desired function is written directly in the memory.

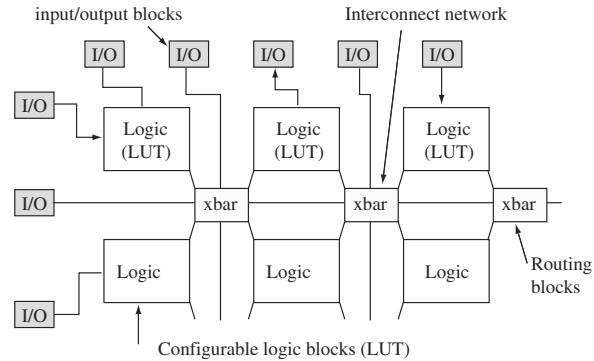
The LUT mechanism is widely used in reconfigurable architectures, which will be discussed later. It has a certain number of very attractive properties in the context of nanotechnology:

- Once one can make memory, one can automatically carry out logic operations. The problem of building operators then reduces to mastering an efficient memory technology.
- The time required to carry out the implanted function does not depend on the operator, but depends purely on the technology and memory structure.
- If the memory technology allows it, the content of the memory can be modified dynamically and the function adapted to suit computation requirements.

However, several more delicate points need to be remembered:

- If the operands are complex or require a high level of accuracy, the amount of memory involved may become prohibitive.
- The apparent simplicity of the LUT idea should not obscure the fact that the system required to address the memory cells can considerably increase the complexity of the whole setup. As shown by feasibility studies [23, 24], memory dots are generally arranged in crossbar-type arrays. Individual addressing of intersections for reading and writing may well require decoding circuits of greater design complexity than the memory dots themselves.

A great deal of work has been carried out on memory. Whatever architecture is envisaged, there are proposals for memory cells. The point is that it is a truly universal function. There could be no computer architecture without memory! Furthermore, as we have just seen, memory can even be self-sufficient



**Fig. 19.9.** Typical structure of a reconfigurable architecture

when it comes to computation. Finally, the fabrication of memory lends itself extremely well to collective production processes and naturally generates highly regular structures. Hence, if the problem is to find efficient addressing techniques, memory devices will certainly figure amongst the first operational realisations of nanotechnology.

### 19.3.2 Reconfigurable Computer Architectures

Reconfigurable computer architectures [26, 27] arise directly from the principles discussed above. Indeed, we have seen that it suffices to modify the content of an LUT to change the implanted logic function. The physical structure of the device remains unchanged whilst its logic structure can evolve as the need arises. Instead of contemplating a straight application of LUT, which would lead to a huge amount of memory, the memory tables used are rather small and distributed within a programmable interconnect network which can connect them together. The combination of simple operators implanted in the LUT units via the communication network can then be used to configure any type of logic circuit. Adjoining input/output facilities to this scheme, one obtains the typical structure of a reconfigurable circuit like the one shown in Fig. 19.9.

A set of memory cells is associated with each resource in the architecture: the LUT for the logic blocks, and a specific memory cell to pilot the state of each switch in the routing blocks. The whole system (logic plus routing) then becomes fully programmable by writing these memories. This type of architecture underlies all currently commercialised reconfigurable circuits, commonly called field programmable gate arrays (FPGA).

Available since the mid-1980s, FPGA circuits are based on standard memory technologies. Up to now, these circuits have only really been used in relatively static applications in which they simply replace conventional logic circuits. In this context, the potential for reconfiguring the circuit is only used for maintenance or updating. However, as soon as these circuits came on the

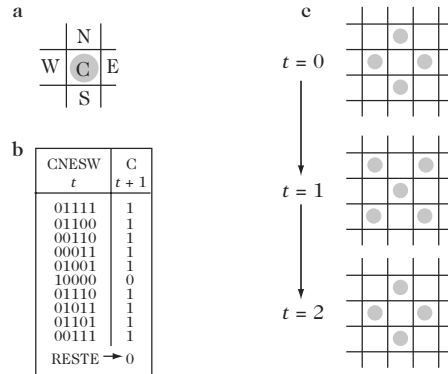
scene, proposals were made to use them at the very heart of a data processing system to design a computational unit whose structure could evolve dynamically depending on the processes to be carried out. This is a very active area of research today in the computer architect community.

One of the main advantages of this architecture lies in the fact that it is essentially based on a memory technology. In addition to this, the array structure of the logic operators within an interconnect network is highly regular. For this reason, the architecture of the reconfigurable processor is perceived by many experts as a natural channel for applying nanotechnology. This idea was first popularised through the HP project known as the Teramac [14]. To begin with the aim of the study was to take advantage of the notion of circuit reconfigurability to enhance the reliability of computers by integrating certain self-repair facilities. The results of the study gave birth to the idea that, if such reconfigurable structures could increase the intrinsic reliability of a system, they might well make it possible to exploit components produced using technologies that were a priori rather unreliable, such as molecular electronics [15]. Moreover, since the only active elements making them up are memory cells and switches, the structures of reconfigurable processors certainly look well-suited to nanotechnological applications.

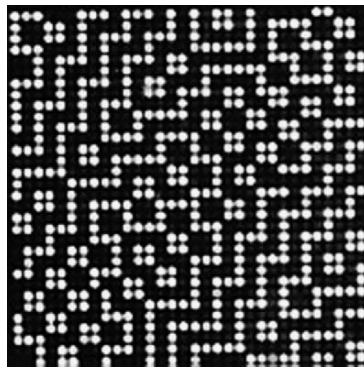
### 19.3.3 Cellular Automata

A cellular automaton comprises a set of very simple units (cells), regularly arranged in a finite dimensional space. Each cell has a state that evolves under the guidance of a systematic rule. The rule, evaluated in discrete time steps, computes a new state  $S_t$  for the cell which depends on the previous state  $S_{t-1}$  of the same cell and the previous states of the neighbouring cells. This neighbourhood generally means only nearest neighbours, i.e., 4 or 8 neighbours for a 2D automaton. Since the rule is a function of a finite set, it is easily specified by a truth table (LUT).

In the example of Fig. 19.10, the chosen rule leads to a cyclical evolution. In fact, the collective behaviour of cellular automata is particularly rich. The complexity of the patterns that can be produced by such simple mechanisms is often striking (see Fig. 19.11). This is one of the reasons why they have been so widely studied, in particular by Wolfram [5] and Toffoli and Margolus [6]. A certain number of applications have thus been proposed to take advantage of this wealth of behaviour. Some rules provide a very good way of generating random numbers. The idea has been suggested for modelling physical phenomena in fluid dynamics and magnetism [6]. Their capacity for massively parallel processing has been assessed in areas such as the routing of messages between resources [7] and image processing [8]. And, of course, we could not discuss the field of cellular automata without mentioning the famous game of Life invented by J. Conway in 1970 [9]. The dynamics created by this rule gives the impression of an artificial life form and the game has stimulated a great deal of interest in cellular automata in general.



**Fig. 19.10.** Two-state 2D cellular automaton. (a) Definition of a neighbourhood with five neighbours (CNESW = centre, north, east, south and west). (b) Definition of a rule giving  $C_{t+1}$  depending on the states CNESW $_t$ . (c) Three steps in the process



**Fig. 19.11.** Pattern generated by a 2D cellular automaton with 9 neighbours and 3 states. Courtesy of the CEA (France)

Cellular automata display especially interesting properties for the production of nanoscale systems:

- In the first place, like any architecture based on the use of memory, they lead to remarkably regular systems, in the sense that all cells are the same and execute the same rule.
- Secondly, as interconnections are limited to nearest neighbours, it is also highly regular.
- The locality of connections is a remarkable feature of cellular automata in the nanotechnological context. It means that short-range interaction modes can be exploited, which were quite unusable in the case of global interconnects. In 1993, researchers at Notre Dame University (USA) suggested a truly original cellular automaton concept based on the electrostatic repulsion of charges in an elementary cell made from 4 quantum

dots [10]. Connecting such cells and exploiting their interactions, one of the two stable states can be propagated along a circuit set up rather like a series of dominos. A whole range of binary logic ‘circuits’ (AND, OR, XOR, INV) has been described on the basis of this idea.

There are nevertheless a few points that can make cell automata difficult to exploit:

- The discrete time synchronous dynamics of the model requires global synchronisation by means of a clock.
- Cellular automata are generally very sensitive to initial conditions and a computer architecture built on this principle requires some way of initialising states. This means that global access to all cells is a prerequisite.

As with memory, collective fabrication techniques might be exploited to produce a very large quantity of cells. However, on the scales discussed in this book, one might expect the level of complexity of the fabricated cells to remain very limited, whence also the complexity of the rules that can be executed. Consequently, one should probably favour an approach which uses the rules of physics to guide the evolution of the nanocells!

#### 19.3.4 Neural Networks

In the computational model inspired by biological neurons, the basic operator is the formal neuron. The model for the formal neuron is already an old concept. Indeed, the first model of the biological neuron is due to the mathematician Pitts and the neurobiologist McCulloch in 1943 [29]. From observations of living cells, they proposed a simplified model of the biological neuron, known as the formal neuron (see Fig. 19.12). The formal neuron thus comprises:

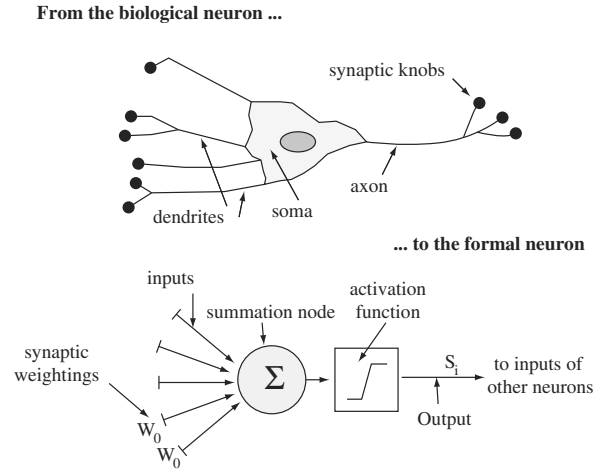
- inputs associated with real weightings which model synapses and dendrites,
- a device calculating the state of the neuron by summing the weighted inputs and applying a nonlinear function which models the cell body (soma),
- an output which communicates the state of the neuron as evaluated previously to the other neurons in the network that models the axon.

Neurons are organised into networks with a wide range of topologies. These fall into two main categories between which all intermediate interconnect schemes are allowed:

- networks arranged in feedforward layers, with forward propagation, for which a simple dynamic scheme is defined by an input  $\rightarrow$  output relation,
- fully connected networks, called recurrent networks, whose many feedback loops endow them with great dynamic richness.

Two distinct process dynamics are applied to govern neural networks built in this way:





**Fig. 19.12.** Modelling the biological neuron

*Relaxation Dynamics.* This calculates the state  $\sigma_i$  of a neuron in terms of the states  $\sigma_j$  of other neurons to which it is connected. This is expressed by

$$\sigma_i = f(H_i), \quad \text{with} \quad H_i = \sum_{j=1}^N \sigma_j W_{ij} + \theta_i. \quad (19.1)$$

In the simplest case, the nonlinear function is a Heaviside step function attributing binary states to the neurons. The variable  $\theta_i$  represents a local weighting (threshold) at the neuron.

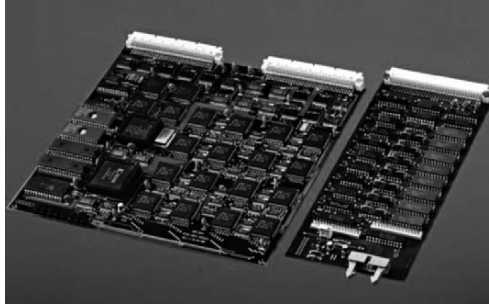
*Learning Dynamics.* This calculates changes in interconnect weightings  $W_{ij}$ , where

$$W_{ij} = W_{ij} + \Delta_{ij}, \quad \text{with} \quad \Delta_{ij} = f(\sigma_i, \sigma_j, H_i, \varepsilon, \dots). \quad (19.2)$$

In contrast to states where the dynamics is well established, there are many algorithms for calculating  $\Delta_{ij}$  from a wide range of parameters and error functions. The simplest learning algorithm is Hebb's rule, which is written  $\Delta_{ij} = \sigma_i \sigma_j$  for signed binary states ( $-1/+1$ ).

However, crude it may seem, this model laid the foundations for a field that has developed considerably since the publication of work by J.J. Hopfield in 1981 [30]. By studying the dynamics of fully connected binary networks, called Hopfield networks, Hopfield showed that under certain conditions ( $W_{ij} = W_{ji}$  and asynchronous dynamics), the network behaves as an excellent energy minimiser. This result complemented earlier work by Rosenblatt [31], who introduced the simplest neural models capable of elementary computation, namely, the perceptron.

This opened the way from formal neural networks to data processing applications, and many neural machines were proposed over the following years,



**Fig. 19.13.** Two examples of neural machines from the 1980s. *Left:* Digital technology. *Right:* Analog technology. Courtesy of the CEA (France)

using all available technological resources, from analog and digital to optical (see Fig. 19.13). Now that nanotechnology has arrived on the scene, these neural architectures may well see a revival of interest. Indeed, many properties of this model seem particularly well suited.

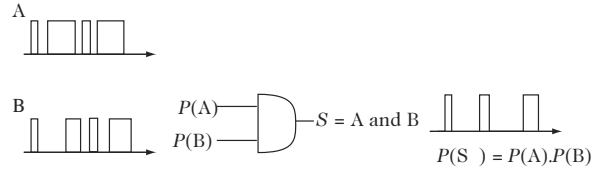
- Neural networks exhibit highly regular architectures. Like cellular automata, all the basic units are identical and carry out the same calculation. This is an eminently parallel architecture.
- The dynamics of neural networks is essentially asynchronous. Theoretically, therefore, there are no problems arising from distribution from a global clock.
- Neural networks are remarkably tolerant of errors. Harmful consequences of fabrication defects or operating errors are reduced, because the operating mode of these architectures results from the collective dynamics of distributed variables.
- As a consequence of its dynamics and the way it is programmed by learning, a neural network can be built from basic components with highly dispersed characteristics.

Possible difficulties are as follows:

- For certain network topologies, the interconnect circuit can become extremely complex, as can control of the associated weighting matrix.
- Direct implantation of neural dynamics as proposed by (19.1) and (19.2) may make it difficult to implement the neuron and synaptic connections. This is due to the realisation of multiplication operators which can be complex and the nonlocality of certain learning algorithms.

In brief, a physical implementation of neural networks depends on the realisation of two critical functions:

- a summing device that can deal with a large amount of input,



**Fig. 19.14.** Two values A and B are encoded in the probabilities  $P(1)$  of two bit sequences. The signals  $P(A)$  and  $P(B)$  are applied at the input of an AND logic gate. The probability  $P(S)$  at the output is the product of the input probabilities  $P(A)$  and  $P(B)$ . The AND logic gate becomes a very simple multiplier

- a synapse function that can realise weighting functions  $W_{ij}\sigma_j$  during the relaxation phase and adaptation functions  $W_{ij} + \Delta_{ij}$  during the learning phase.

A certain number of ideas have been published concerning synapses made using molecular single-electron transistors [32] and adaptive synapses [33] capable of locally evaluating a learning rule like the Hebb rule. Although this work remains rather conceptual, one can expect experimental demonstrations of neural nanodevices to follow soon enough, given the attractive aspect of the neural approach in the context of nanotechnology.

## 19.4 Computer Environment

### 19.4.1 Information Coding

The binary code underlies the coding used in computers. If we consider the circuits in Fig. 19.5, we observe that the layout of the transistors is a direct consequence of the way information is coded for the NAND and NOT functions. If the same functions had been built using some code other than the natural binary coding, the architecture would have been totally different.

As an illustration, Fig. 19.14 shows a multiplier function for probabilities. The data is very simply encoded in the probability of having a state 1 on a signal consisting of a bit sequence. Note the simplicity with which the multiplication is accomplished. Although such coding is obviously very limited in the context of a numerical computation to arbitrary accuracy, it proves extremely efficient for more qualitative operations.

### 19.4.2 Defect Tolerance

Current microelectronics systems integrate many internal and external devices to ensure smooth running and periodic test phases. Generally speaking, as a system grows more complex, the relative part devoted to control and test functions also grows. This is particularly true for the latest von Neumann

computer architectures, which integrate many computation units and cache memory levels. Indeed, for this type of architecture, tolerance of fabrication defects is very low. As many functions must therefore be tested as possible if one hopes to optimise research and development investments.

The use of nanotechnology in computer design requires some revision of previously well-established concepts. Indeed, most technologies yet envisaged, e.g., molecular electronics, carbon nanotubes, do not have the same properties as the transistor technologies so successfully used up until now. To begin with, the small size of these devices makes it impossible to exercise perfect control over their fabrication. Secondly, at length scales below 10 nm, quantisation effects can no longer be neglected. Finally, electron transport in very small conductors can seriously perturb the quality of transmitted information.

In order to overcome these difficulties, there are two main channels:

- The first consists in trying to correct or minimise induced errors by self-correction or redundancy systems, in order to continue using known computer architectures, by making the underlying devices more reliable.
- The second consists in trying to bypass the stochastic nature of the devices, conceiving of new architectures and computation techniques for which the overall result is deterministic even though the elementary devices are not.

Table 19.3 summarises the properties of computer architectures discussed here. Apart from the physical properties used to assess the suitability of a solution for a given technological target, the table also shows properties relevant to the use of the architecture. Indeed, what could be done with a perfectly made machine if there were no way of exploiting it? This is why it is important to examine all the implications of a given architecture for the computer system as a whole.

A multidisciplinary approach is essential to achieve this. Indeed, the processing system must be viewed as a whole in order to arrive at conclusive results. A mastery of the elementary device does not guarantee its optimal use within a computer system, and even less that it will contribute effectively to the accomplishment of some useful task. Moreover, when one considers that nanotechnological devices will necessarily be integrated into more complex, technologically heterogeneous systems, the design of interfaces becomes the central problem. It is only by understanding the system as a whole that such hybrid computational tools will become effectively exploitable. Considerable interaction between technological specialists and computer architects will thus be required at all development stages. For example, it is striking to observe that early work in this area sought to obtain nanodevices that would directly achieve Boolean logic operations. However, it turns out that, on the nanometric scale, the noisy nature of phenomena can make this a delicate task. Moreover, a Boolean logic operator is not necessarily the most crucial function in current computer architectures.

Although at first glance it may seem relevant to carry out such work for the purpose of direct comparison between nanotechnological products and

those arising from conventional technology, this strategy may actually hinder the emergence of new technologies. Indeed, points of comparison provided by speed, passband, power, etc., are too strongly biased toward the technology that produced them. Continued exchange between specialists from the relevant fields gives a faster way of guiding work into genuinely useful channels.

Consequently, it is very important for the development of future computer generations to face the problem of designing heterogeneous machines. Indeed, even though nanotechnology is likely to succeed in stimulating the appearance of new data processing architectures like the ones we have been discussing, no one believes that the latter will completely replace computing machines as we know them today. To convince the reader, it is enough to examine a little more closely the fields of application for which each of these proposals is potentially best suited in order to realise that there is actually no direct competition, but instead, a great deal of complementarity. Today's processors are first-rate calculators, unbeatable experts when it comes to arithmetic. All tasks concerned with pure numerical computation are likely to remain within their jurisdiction as long as precise results are required. But real applications are made up of a wealth of elementary tasks which do not all require absolute numerical precision. For example, in the case of an operation which compares two values, what matters is the order of magnitude rather than the exact value. For another example, consider a task in which one evaluates the respective 'weights' of a set of quantities, to weigh up the pros and cons so to speak, in order to reach a compromise or make a decision. In this situation, the algorithm designer will have recourse to a whole range of numerical tricks with which to achieve evaluation and comparison functions. These two types of processing are thus classified according to whether they are quantitative (numerical values, high-precision calculation) or qualitative (low accuracy, compromise). Here one can imagine a possible division between the types of architecture depending on the type of processing – quantitative or qualitative – for which they are best suited.

As soon as we consider the possibility of pooling together several data processing resources, sometimes based on different technologies, two key questions must be addressed:

- Are the technologies compatible enough to integrate the parts into a single whole? Can the different fabrication processes be used together to build the multiple resources without each somehow altering the characteristics of the others?
- If we succeed in fabricating the ensemble, will we be able to set up connections that allow the required dialogue between resources, in particular on the data encoding level?

Several examples already show that essentially very different technologies can be combined, such as MRAMs with CMOS technology. Real progress has also been made in the deposition, and even in situ growth, of nanowires and carbon nanotubes. However, whichever of the architectural proposals made here is

**Table 19.3.** Properties of several computer architectures

	Computation unit	Temporal dynamics	Connections	Defect tolerance	Standard use	Field of application
Von Neumann computer	Complex, irregular	Synchronous	Global, global memory	Low (corrections)	Central sequential program	Numerical calculation and broad spectrum control (quantitative)
Reconfigurable architectures	Rather simple, regular	Mixed	Mixed	Quite good (reconfiguration)	circuit combination, truth tables	Numerical calculation, data flow (quantitative)
Cellular automata	Simple, highly regular	Synchronous	Local, distributed memory	Average (depending on rule)	Rule for evolution, parallel	Fine-grained regular data
Neural networks	Simple, regular	Asynchronous	Depends on, topology, distributed memory	Good (intrinsic)	Learning, parallel	Approximation of functions, optimisation (qualitative)

actually put into practice, if any, the successful development of symbiotic computers will hinge upon the way the interfaces and the system as a whole are handled.

## 19.5 Prospects

The different types of architecture discussed here are very varied in terms of both structure and potential use. They can be distinguished in particular by the complexity of the elementary computation unit and the interconnect network. But they can also be distinguished in terms of their ability to help with solving non-trivial problems. Although computer architectures based on von Neumann processors have already proved their worth, the same cannot be said for some of the architectures we have been discussing here. In particular, the cellular automata which represent such an attractive channel for nanotechnology due to their great regularity have not yet been provided with a well-established programming plan, or one might rather say, a proven guide to their correct use. This does not mean that one should abandon these architectures, but it should be borne in mind that, quite apart from the problem of physical implementation (the main concern of this book), these new architectural ideas also require a great deal of research upstream concerning programming paradigms and data encoding problems.

The reader will have noticed the omission of several rather original proposals. One example is the idea of a computer based on DNA, and several variants, proposed by Adleman [16] in 1994 to solve problems of combinatoric optimisation. Another is the quantum computer arising from ideas originally due to Feynman [17] and later developed by Deutsch [18]. Although in the first example (DNA), many implementation problems await solution, in the second (quantum computing), several algorithms of some interest have been devised [19]. As a result, much work is under way in the latter field, not all of it related to nanotechnology. Having said this, neither of these proposals gives reason to expect concrete developments in the near future. This is why we have concentrated mainly on architectural solutions that can easily be extrapolated in the mid-term from machine designs that are already well tested.

So we must leave it to the future and to the various parameters that determine it – not all very scientific – to tell us what components and what architecture will be used to build the computer of tomorrow.

## References

1. Davis, M.: *The Universal Computer: The Road from Leibniz to Turing*, W.W. Norton & Company (2000)
2. Aspray, W., et al.: *Computing Before Computers*, IOWA State University Press, Ames (1990)
3. Moore, G.: Cramming more components onto integrated circuits, *Electronics* **38**, 8 (1965)
4. Aspray, W.: *John Von Neumann and the Origins of Modern Computing*, MIT Press (1990)
5. Wolfram, S., et al.: *Theory and Applications of Cellular Automata*, World Scientific (1986)
6. Toffoli, T., and Margolus, N.: *Cellular Automata Machines: A New Environment for Modeling*, MIT Press (1987)
7. Ryan, T., and Rogers, E.: *An ISMA Lee Router Accelerator*, IEEE Design and Test of Computers, pp.38–45 (October 1987)
8. Popovici, A., and Popovici, D.: *Cellular Automata in Image Processing*, Proceedings of MTNS 2002, Notre Dame University (2002)
9. Gardner, M.: The fantastic combinations of John Conway’s new solitaire game ‘life’, *Scientific American* **223**, 120–123 (1970)
10. Lent, C.S., Tougaw, P.D., Porod, W., Bernstein, G.H.: *Nanotechnology* **4** (1), 49–57 (1993)
11. Lloyd, S.: *Phys. Rev. Lett.* **88**, 23 (2002)
12. Butts, M., DeHon, A., Goldstein, S.: *Molecular electronics: Devices, Systems and Tools for Gigagate, Gigabit Chips*, Proceedings of ICCAD (2002)
13. Goldstein, S., and Budiu, M.: *NanoFabrics: Spatial Computing Using Molecular Electronics*, Proceedings ISCA (June 2001)
14. Anderson, R., Carter, R., Cullbertson, W.B., Kuekes, P., Snider, G.: *Teramac – Configurable Custom Computing*, Proc. of IEEE FCCM’95 (1995) pp. 32–38
15. Heat, J., Kuekes, P., Snider, G., Williams, S.: *Science* **280**, 5370 (1998)
16. Adleman, L.: *Science* **266** (November 1994)
17. Feynman, R.: Simulating physics with computers, *Int. J. Th. Phys.* **21** (6/7), 467–488 (1982)
18. Deutsch, D.: Quantum theory, the Church–Turing principle and the universal quantum computer, *Proc. Roy. Soc. Lond. A* **400**, 97–117 (1985)
19. Shor, P.W.: *Algorithms for Quantum Computation: Discrete Log and Factoring*, Proc. of 35th IEEE FOCS (1994) pp. 124–134
20. Hey, A.J.G. (Ed.): *Feynman and Computation: Exploring the Limits of Computers*, Perseus Books (1999)
21. Rumelhart, D., McClelland, J. (Eds.): *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, MIT Press (1986)
22. Likharev, K.K.: In: *Nano and Giga Challenges in Microelectronics*, ed. by J. Greer et al. (Elsevier, Amsterdam, 2003) pp. 27–68
23. Lieber, C., and Cui, Y.: *Science* **291** (February 2001)
24. Lieber, C., et al.: *Science* **289** (July 2000)
25. Wirthlin, M.J., Gilson, K.L., and Hutchings, B.L.: *A Dynamic Instruction Set Computer*, Proceedings IEEE Workshop on FPGA for Custom Computing Machines, Los Alamos (April 1995)
26. Gray, J., and Kean, T.A.: *Configurable Hardware: New Paradigm for Computation*, Decennial Caltech Conference on VLSI (March 1989) pp. 277–293



27. Bertin, P., Roncin, D., Vuillemin, J.: Introduction to programmable active memories. In: *Systolic Array Processors*, Prentice Hall (1989) pp. 300–309
28. ITRS Roadmap for the Semiconductor Industry (2003)
29. McCulloch, W.S., and Pitts, W.: *Bull. Math. Biophysics* **5**, 115–133 (1943)
30. Hopfield, J.J.: *Proc. Natl. Acad. Sci. USA* **79**, 2554–2558 (1982)
31. Rosenblatt, F.: *Psychological Review* **65**, 386–408 (1958)
32. Türel, O., Lee, J.H., Ma, X., Likharev, K.K.: *Int. J. of Circuit Theory and Applications* **32**, 277–302 (2004)
33. Rouw, E., and Hoekstra, J.: *Bio-Inspired Stochastic Neural Networks for Nanoelectronics*, Proc. CASYS 2001, 5th International Conference (2001) pp. 501–513
34. Javey, A., Kim, H., Brink, M., Wang, Q., Ural, A., Guo, J., McIntyre, P., McEuen, P., Lundstrom, M., Dai H.: *Nature Materials* **1**, 241–246 (December 2002)
35. Javey, A., Guo, J., Wang, Q., Lundstrom, M., Dai, H.: *Nature* **424**, 654–657 (August 2003)
36. Keren, K., Berman, R.S., Buchstab, E., Sivan, U., Braun, E.: *Nature* **424**, 1380–1382 (2003)

---

## Index

- ab initio methods, 183, 305, 306, 457, 759, 763, 765–773
  - GW approximation, 762
  - local density approximation, 766, 769–770
  - self-consistency, 770–771
- Abbe criterion, 619
- acoustic phonon, 33
- ADP, 741
- adsorbate, 42
  - periodic growth, 42–53
- AFM, *see* atomic force microscope
- Airy disk, 714
- aliasing, 535
- alkanethiol, 372, 373
- alloy, 55, 109, 334
  - copper–nickel, 539
  - FePd, 588, 591, 616
  - FePt, 588, 616
  - ferromagnetic, 510
  - for STM tip, 71
  - GeSbTe, 583
  - IV–IV, 409
  - L1<sub>0</sub>, 616
  - magnetic, 260
  - magnetic anisotropy, 517
  - phase-change, 594, 595, 609
  - SiGe, 409
  - surface, 51
  - tellurium, 595
- aluminium, 342
  - cluster, 214, 215
  - film, 340
  - interconnects, 399
  - aluminium oxide matrix, 335
- AM-AFM, 101
- amphiphile, 369, 375, 472, 490, 491
  - hydrophobic tail, 369
  - polar head, 369
  - self-assembly, 370, 371
- annealing, 59, 60, 63, 248, 375, 388, 485
- annihilation operator, 338
- anti-bonding state, 190, 270
- antibody/antigen recognition, 669, 687, 688
- antiferromagnetic interaction, 507, 532, 536, 537, 588–590, 617
- antiferromagnetism, 508, 509, 608
  - of bulk chromium, 539
- aperture SNOM, 133–136
- apertureless SNOM, 131–133
- argon, 12
  - cluster, 189
- Arrhenius law, 243, 547, 587
- artificial atom, 641
- artificial muscle, 368
- aspect ratio, 160, 161
- atomic chain, 342
- atomic force microscope, 90–118, 123, 133, 136, 173, 594
  - amplitude modulated, 101
  - applications, 115–118
  - approach–retract curve, 104
  - as nanoindenter, 112
  - atomic resolution, 107
  - cantilever, 91–93

- contact mode, 94, 98–101
- diamond tip, 112
- electromagnetic measurement, 109–111
- finite size effects, 96
- for biology, 742
- force curve, 98–100, 107, 108, 112, 113
- force measurement, 107–109
- frequency modulated, 101
- friction mode, 94, 100–101
- imaging modes, 92–95
- in lithography, 20
- linear resonant mode, 102–103
- manipulation, 115
- phase image, 105
- resolution, 95–98
- resonant mode, 95, 101–107
- setup, 91–92
- shear-force mode, 132
- tapping mode, 95, 103–107, 132
- tip, 96, 97, 317
- tip apex, 105
- tip–sample interaction, 102, 104
- topography, 105, 106
- ATP, 727
  - hydrolysis, 741
- Auger effect, 639
- autocorrelation function, 716
- avalanche photodiode, 709
- Aviram–Ratner model, 472
  
- back-gate configuration, 479
- ballistic regime, 336, 419
- band
  - diagram, 81, 483, 657
  - gap, 200, 261, 270, 409, 640, 762
  - structure, 207, 758, 762, 765
- Bardeen approximation, 73
- BCS mechanism, 216
- beam spreading, 31, 32
- benzene, 469, 709
  - absorption spectrum, 771
  - dithiol, 470
- binary coding, 782
- biological medium, 669, 670, 698, 724, 735–737
- biological vector, 262, 370
- biology, 670
  
- biomimetic material, 691
- biomolecule, 448, 686, 688–691, 697
  - single, 737–743
- biophotochip, 669
- biophotonics, 682, 683, 686, 698–707
- biosensor, 165, 624–626, 669
- biotin, 687, 691
- blinking, 698, 715, 743
- Bloch
  - function, 293, 637, 647
  - mode laser, 654
  - theorem, 292
  - wall, 521, 523, 524
- block copolymer, 105
- Bohm–Aharonov effect, 437–439
- Bohr–Sommerfeld theory, 429
- Boltzmann
  - distribution, 755
  - transport, 417
- bond cutoff model, 183, 184
- bonded film, 59, 60
- bonding state, 190, 270
- Boolean logic, 782, 783
- Born–Oppenheimer approximation, 678, 772
- boron nitride, 301
  - nanotube, 301–303
- bottom-up approach, 41, 241, 326, 332–335, 349, 674, 688
- Bragg mirror, 79, 647–648
- break junction, 336, 339, 340, 342, 451, 469–471, 475
- Brillouin zone, 648, 762
  - first, 183, 294, 647
  - graphene, 295
  - K point, 655
- Brownian motion, 727, 729
- Bruggeman model, 264
- buckminsterfullerene, 281
- buffer cell, 442, 443
- Burger circuit, 44
- buried dislocation, 59
  
- C<sub>60</sub> molecule, 87, 115, 189, 191, 192, 261, 280, 282, 285–288, 313, 475, 477
  - chemical properties, 287–288
  - cycloaddition reactions, 289
  - cyclopropanation, 288

- electrochemical properties, 287
- five-level model, 285, 286
- ISA, 286
- nonlinear absorption, 285
- photophysical properties, 285–287
- solubility, 285
- structure, 283
- transport properties, 454–456
- Camley–Barnas model, 565–567
- capillary electrophoresis, 165, 168
- capillary moulding, 328
- carbon nanotube, 110, 111, 115, 279–318, 449, 795, 796
  - applications, 313–318
  - armchair, 290, 291, 295–297, 753
  - as probe, 316
  - assembly, 492
  - atomic structure, 280, 319
  - chemical properties, 313
  - chemistry, 317
  - chiral, 290, 291, 298–300, 311, 753
  - chiral angle, 291, 309, 315
  - chiral vector, 291
  - components, 479–489
  - conductance, 312
  - crystal structure, 289–291
  - deposition, 494–496
  - discovery, 281–282, 450
  - doped, 313
  - ductility, 755
  - electron emission, 316
  - electronic structure, 295–300, 319
  - Fermi level, 313
  - filled, 313
  - functionalisation, 317, 318
  - growth, 305–307, 493
  - hardness, 755
  - helicity, *see* carbon nanotube, chiral angle
  - interconnects, 314, 785
  - large scale production, 304, 305, 315
  - matrix memory, 613
  - mechanical properties, 312
  - metallic, 297–299, 312, 487–489
  - MWNT, 301, 303, 305, 308, 312, 316, 317, 753
  - network, 497
  - observation, 308–311
  - properties, 311–313
  - roll-up vector, 291
  - rope network, 496
  - self-assembly, 300, 305, 315
  - semiconducting, 298–300, 314, 479, 494, 785
  - SET, 486–489
  - sorting, 494
  - stressed, 755, 756
  - switch, 497
  - SWNT, 301, 303–306, 309, 310, 479, 753
  - synthesis, 302–305
  - technology, 785
  - transistors, 785
  - zigzag, 290, 291, 297–298, 753
- carborane, 364
- carotenoids, 469
- CARS, 714, 736, 737
- catenane, 355, 366, 476, 478
  - doubly-interlocked, 356
- catenate, 354, 356
  - trefoil knot, 355, 356
- cavity mode, 627
- CD, 583, 584, 634
- CD-R, 583
- CD-RW, 583, 595
- cell engineering, 165
- cellular automata, 498, 789–791, 797, 798
  - game of Life, 789
- centrosymmetry breaking, 701, 704, 705, 735
- charge transfer interaction, 354, 365, 367
- chemical vapour deposition, 54, 55, 304, 306, 315, 493
- cholesterol oxidase, 720, 721
- chromatic aberration, 32
- chromium, 539
  - film, 14
- chromophore, 674, 720
- chymotrypsin, 725, 727
- circuit speed, 383
- citrate, 247, 248, 253, 276, 689
  - as stabiliser, 251
- clathrate compound, 191
- Clausius–Mossotti polarisability, 153
- closing defect, 189
- cluster, 49, 57, 179–277, 750

- absorption cross-section, 229, 233, 235
- absorption spectrum, 231, 238, 264, 771
- aluminium, 214, 215
- array, 262–263
- assemblies, 252–266
- beam, 254–256
- bimetallic, 261
- cage, 261
- chemical reactivity, 750
- circular, 191
- coalescence, 258–259, 754
- cobalt, 250
- collective excitations, 225–241
- core–shell, 240, 248, 261, 696
- covalent, 200, 261, 270–271
- covalent binding, 190–191
- cubic, 249
- cubo-octahedral, 182
- deposition, 254–256
- diffusion, 257–259
- divalent metal, 199
- dynamic polarisability, 228, 773
- dynamics, 755
- electron shell structure, 207–217, 221–225
- electron supershell structure, 217–225
- energy levels, 204
- equilibrium shape, 180–193
- extinction spectrum, 265
- facetted, 249
- finite-size effects, 206–241
- fluctuations, 200–205, 215
- functionalised, 251, 254, 262
- gallium, 220, 221, 223
- gold, 193, 195, 238–240, 251, 253, 254, 256–258, 262, 263, 276
- heat capacity, 205
- in matrix, 750
- ionic binding, 192–193
- ionisation potential, 196–197
- island, 256–259, 262
- Lennard-Jones binding, 259
- lithium, 231
- magic, 211, 212, 220, 221, 232, 273
- magnesium, 203
- magnetic, 260–261, 531
- mass spectrum, 210, 212–215, 220, 223
- melting temperature, 193–196, 200–202, 215
- memory effect, 230, 258
- metallic, 184, 196–200, 209, 213, 216, 225, 227, 232, 234, 264
- metallic binding, 184–189
- metastable, 248
- molybdenum, 255
- monovalent, 198
- multilayer, 248
- noble metal, 237, 238
- nucleation, 241–246
- optical properties, 225–241
- palladium, 453
- paramagnetic susceptibility, 202–205
- parity, 203
- photoionisation spectrum, 214
- potassium, 231
- preparation, 241–251
- radius, 193–200
- silicon, 261
- silver, 205, 238–240, 249, 265
- silver sulfide, 250
- size, 258, 269
- sodium, 209, 212, 215, 219–221, 233, 234
- soft ionisation, 214
- spherical, 191, 248
- stabilisation, 251
- static polarisability, 228, 230, 231, 234, 236
- transition metal, 199, 261, 766
- van der Waals binding, 189, 259
- CMOS, 383, 385–391, 497
- inverter, 386–388, 390–391
- logic, 390
- technology, 783, 796
- transistor, 12
- cobalt, 56, 261, 510, 515–517, 561, 586, 588
  - as stabiliser, 305
  - cluster, 182, 183, 250, 547
  - dots, 57, 58, 163
  - film, 526
  - islands, 56, 531
  - nanoparticle, 373, 547
  - nanostructure, 58

- ultrathin film, 526, 540
- coercive field, 588, 591, 608, 615
- coherence length, 193, 418, 433, 434, 437
- coherent anti-Stokes Raman scattering, *see* CARS
- cohesive energy, 180–181, 184, 186, 267
- colloid, 179–277
  - assemblies, 252–266
  - metal, 246–251
- colloidal suspension, 371
- complexation, 354
  - axial, 357, 358
- computer architecture, 776–798
  - 2-bit adder, 782, 783
  - binary code, 794
  - bistable circuit, 783
  - defect tolerance, 794–798
  - FPGA, 788
  - interconnects, 780, 782–783
  - memory, 779–781
  - NAND operator, 783, 794
  - new ideas, 786–794
  - NOT operator, 783, 794
  - operator, 780, 782–783
  - parallel processing, 789, 793
  - processing unit, 779, 780, 785
  - reconfigurable, 788–789, 797
  - register, 780, 781
  - von Neumann, 779, 780, 785, 795, 797
- conductance, 335, 418, 431, 434
  - drain, 385, 390–393
  - Drude, 432
  - electron gas, 432
  - nanotube transistor channel, 485
  - of carbon nanotube, 312
  - of molecule, 374
  - of nanowire, 335–343
  - quantisation, 336, 432
  - SET, 487
  - single-molecule, 452, 453, 457, 469, 470
  - universal fluctuations, 435–436
- conducting multilayer, 532
- confinement model, 271
- confocal microscopy, 676, 691, 706, 709, 711–715
- conformation, 681
  - dynamics, 719, 724–726
- coordination number, 56, 57, 183, 184, 186, 187, 257, 269
- copper, 45, 237, 529, 538, 539
  - Fermi surface, 537, 538
  - interconnects, 399
  - surface, 56, 57, 86, 331
- copper–oxygen stripes, 47
- core–shell cluster, 240, 248, 261, 696
- Coulomb
  - blockade, 342, 420–425, 427, 428, 454, 464, 486, 487, 489, 612, 642
  - oscillations, 421–422, 428
  - regime, 458, 465–469
  - stairs, 454
- CQED, 644, 646
- creation operator, 338
- cross-linking agent, 688, 689, 691
- crown ether, 707
- cryotron, 440
- crystal
  - bcc, 516
  - cubic, 516
  - defect, 543
  - fcc, 516
  - growth, 334
  - ionic, 621
  - lattice, 508, 515, 517, 518, 529, 535, 576, 637, 646
  - structure, 751
  - surface, 42–47
- cube, 184, 185
- cubo-octahedron, 183, 185, 266
- Curie
  - susceptibility, 205
  - temperature, 508, 617
- cyclodextrin, 363, 364
- cyclophane, 363
- cytidine, 359
- cytosine, 359
- dangling bond, 4, 190, 191, 200, 271
  - silicon, 87
- data processing, 777, 778
- data storage, 417, 443, 503, 583–618
  - antiferromagnetically coupling media, 588–590
  - bit, 548, 587, 591, 595–596
  - discrete media, 592–593
  - error rate, 547–548, 599

- GMR read head, 585–586
- grain boundary, 587, 589
- heat-assisted recording, 591
- local probe techniques, 584, 593–595
- longitudinal medium, 585
- magnetic grain, 586, 587, 592
- nanoscale bit, 585
- perpendicularly magnetised media, 590–591
- phase-change material, 595, 609
- retention time, 584, 597, 600, 603, 616, 780
- switching time, 587
- word, 595–596
- write time, 618
- DDA model, 257
- de Broglie
  - relation, 761, 762
  - wavelength, 418, 420
- decision-making cell, 442, 443
- decoherence, 576
- defect, 656–661
  - as cavity, 657–660
  - as wave guide, 660–661
- delocalised electronic structure, 78
- demagnetising field, 512–515, 543, 576, 589, 591, 615
- dendrimer, 349–353, 364, 720
  - convergent synthesis, 352–353
  - divergent synthesis, 350–352
  - monodispersed, 352
- dendron, 352, 353
- density functional theory, 208, 472, 474, 767–769
  - time-dependent, 771
- density of states, 218, 219, 424, 474, 511, 538, 557, 560, 561
  - of single QD, 640
  - semiclassical theory, 221–225
- depolarisation factor, 622
- developer, 6, 8
- dextran, 108
- DFT, *see* density functional theory
- diamond, 191, 280–281
  - AFM tip, 112
  - film, 112, 113
  - structure, 190
- diblock copolymer, 371, 374, 375
- diffraction, 627
  - by two slits, 141
  - grating, 121, 624, 626, 628, 629
  - limit, 19, 20, 24, 137–142
- diffusion, 433, 726–731
- diffusive regime, 336
- dimerisation, 343
- diode
  - current rectifying, 471
  - laser, 659
  - light-emitting, 639, 659, 674
  - molecular, 449, 471–474
  - NDR, 472
- dipole approximation, 226, 228
- dipole radiation, 143–144, 147–155
  - near an object, 149–150
  - near nanoparticle, 152–155
  - near plane mirror, 151–152
- dislocation, 639, 640
  - network, 45
- distributed feedback laser, 631, 654
- DLVO model, 108
- DNA, 105, 165, 168, 172, 251, 353, 449, 453, 687, 688, 690, 695, 720, 738
  - computer, 798
  - tile, 498
  - transcription, 740, 742
- dodecahedron, 185, 188
- dodecanethiol, 250
- DRAM, 584, 596, 604–605, 611, 781
  - 1T-capacitorless, 609–610
  - bit line, 604
  - retention time, 604
  - storage capacity, 605
  - word line, 604
- Drude
  - conductance, 432
  - metal, 621
  - transport theory, 429
- Drude–Sommerfeld model, 228, 232, 233
- dry etching, 11–13
- DVD, 583, 584, 595, 609
- dye, 370, 695, 727, 736, 737
- dysprosium silicide nanowire, 334
- Eccles–Jordan flip-flop circuit, 783
- EELS, 308
- EEPROM, 596, 601, 611
  - flash, 601, 602

- EFISH, 701
- elastic energy, 52, 53
- elastic stress engineering, 58
- elastomer, 112
  - viscoelastic phase, 113
- electric force microscopy, 110
- electrodeposition, 158, 162, 453
- electroluminescent diode, 375
- electrolytic growth transfer, 15–18
- electron
  - back-scattered, 31, 32, 34
  - coherence length, 418, 433, 434, 437
  - coherent transport, 456
  - confinement, 79, 532–540, 761
  - delocalisation, 269, 283
  - delocalised, 207, 209, 256, 264
  - effective mass, 760
  - elastically bound, 147
  - energy levels, 209
  - Fermi energy, 418
  - gas, 761
  - heat capacity, 205
  - incoherent transport, 342
  - indistinguishability, 507
  - magnetic moment, 506
  - magnetic susceptibility, 203
  - mean free path, 418, 419, 554, 563, 567, 569
  - mobility, 386, 409, 410
  - numerical simulations, 757–773
  - phase coherence length, 336
  - quantum theory, 757–759
  - spillover, 234–237, 240, 264
  - spin, 203, 506
  - spin diffusion length, 554, 563, 567, 569
  - spin-dependent transport, 557, 559–568
  - transport, 336, 429, 450, 454–470, 511, 773, 795
  - wave function, 69, 72–74, 417–419, 428, 434, 533, 637, 758, 761, 772, 773
  - wavelength, 28
  - weak localisation, 432–434
- electron beam lithography, *see* lithography, electron beam
- electron microcolumn array, 29
- electron microscope, 30
- electron–electron interaction, 341, 343, 759, 771
- electron–hole interaction, 270
- electron–ion interaction, 223, 225, 237, 759, 769, 771
- electron–magnon interaction, 560, 562
- electron–matter interaction, 31
- electron–phonon interaction, 80, 341, 343, 766
- Encyclopedia Universalis, 87
- endoreceptor, 365
- entangled photons, 628, 644
- enzyme, 690, 720, 726
  - detector, 489
- epitaxy, 15, 17, 257, 332, 591, 636, 639, 640
  - coherent, 50
  - condition, 257, 259
  - incoherent, 50
  - liquid phase, 55
  - molecular beam, 53, 60
  - vapour phase, 55
- EPRM, 596, 601
- ergodic theorem, 436, 685
- etch mask, *see* lithography, mask
- Euler angles, 676, 677, 685
- Euler theorem, 191, 282
- evanescent wave, 122, 124–127, 137, 139, 142–143, 623, 624, 631, 651, 667, 711
- exchange
  - interaction, 507, 514, 520–522, 540, 552, 559
  - length, 521–523, 540
- excimer laser, 25, 168, 406
- excitation transfer, 724–726
- exciton, 270, 677, 695, 766
- exoreceptor, 365
- exposure time, *see* lithography, dose
- Fabry–Perot device, 532, 657
- factory roof structure, 555
- Faraday law, 17
- fcc lattice, 45, 46, 183, 187, 190
- FeRAM, 596, 606–607
  - bit line, 607
- Fermi
  - distribution, 337, 535
  - gas, 758



- golden rule, 150, 151, 422–424, 467, 681, 683
- level, 72, 74, 83, 197, 198, 202, 211, 219, 237, 299, 300, 421, 458, 485, 534, 557, 560, 610, 611
- surface, 535–537, 539
- wavelength, 234, 240, 264, 336, 418, 419, 430, 433
- Fermi–Dirac distribution, 202, 273, 423, 565, 635
- fermion, 207, 273
- ferrimagnetism, 509
- ferrofluid, 371
- ferromagnetic
  - film, 523, 526
  - interaction, 507, 532, 536, 537
  - metals, 510–511, 517, 536, 542, 559
  - semiconductor, 559
- ferromagnetism, 508, 509, 514
  - itinerant, 530
- Fick law, 728
- finite-size effects, 206–241
- FISH, 694, 697
- flavin adenine dinucleotide, 720
- flocculation, 250, 688
- fluctuations in nanosystems, 200–205
- fluorescein isothiocyanate, 709
- fluorescence, 123, 253, 669, 697, 700
  - as biological marker, 691–695
  - intensity, 682
  - lifetime, 710, 725
  - microscopy, 711
  - multiphoton, 681
  - one-photon, 681, 682, 711
  - quantum yield, 693, 705, 706, 708, 725
  - single-molecule, 137
  - two-photon, 682, 732–734
- fluorescence correlation spectroscopy, 716–718
- fluorescent protein
  - green, 673, 720
  - red, 673
- fluoroionophore, 700
- fluorophore, 251, 709, 720, 730, 734
- fluxon, *see* magnetic flux quantum
- FM-AFM, 101
- force feedback nanomanipulator, 117
- four-wave mixing, 705
- Fowler–Nordheim effect, 602
- FPGA, 788
- fractal, 253
- fractional charge transfer, 457
- FRAM, 584
- Franck–Condon approximation, 678
- Frank–van der Merwe growth, 49
- Fraunhofer diffraction, 20, 24
- free-electron model, 341
- Fresnel diffraction, 20
- FRET, 678, 724, 725, 727
- friction, 114
- friction force microscopy, 94
- fuel cell, 317
- fullerene, 261, 279–318, 370
  - C<sub>60</sub>, *see* C<sub>60</sub> molecule
  - C<sub>70</sub>, 283
  - discovery, 281
  - doped, 752, 769
  - isomers, 282, 283
  - production, 284
  - smallest, 282
  - stability, 283
  - structure, 282–284
  - stuffed, 191
- GaAs, 79, 80, 261, 638, 639, 652, 659
  - band gap, 82
  - cavity, 641
  - components, 408
  - mesa, 37
  - nanowire, 334
  - photonic crystal, 164
  - substrate, 58, 59, 640
  - wire, 434–436
- gallium cluster, 220, 221, 223
- GaN, 640
- gas sensor, 489
- Gaussian
  - beam, 34, 35
  - probe, 30
- Ge quantum dots, 60
- gene, 353
- genetic engineering, 370
- geometrical optics, 24
- giant atom, 213
- giant magnetoresistance, *see* magnetoresistance, giant
- Gibbs pressure, 181, 194

- Gibbs–Duhem relation, 193  
 Gibbs–Thomson effect, 334  
 GILD, 406  
 glass surface, 134, 170, 372  
 glass transition temperature, 158  
 gold, 33, 45, 115, 196, 237, 330, 336, 538, 624, 697  
   cluster, 193, 195, 238–240, 251, 253, 254, 256–258, 262, 263, 276  
   dielectric function, 238  
   electrode, 111, 469, 470  
   film, 170, 340  
   icosahedron, 190  
   islands, 56, 57, 63, 64  
   loop, 438  
   nanoparticle, 134, 689  
   nanostructure, 57, 63, 116, 117, 691  
   nanowire, 333, 343, 375  
   on AFM tip, 115  
   particle, 632, 633  
   pillars, 18  
   reconstruction, 43, 44, 56, 58  
   shell, 248  
   silver, 238  
   substrate, 173, 257, 258, 454, 456  
   surface, 45, 56, 328, 330, 372, 373, 688  
   vicinal surface, 48, 57  
   wire, 454  
 golden section, 187  
 GPS, 409  
 grain boundary, 435  
 graphene, 289–291, 319  
   electronic structure, 292–295  
   first Brillouin zone, 295, 296, 298, 299  
   K point, 298, 299  
   lattice, 292, 293  
   reciprocal lattice, 295  
   transport properties, 293  
 graphite, 113, 280–281, 333  
   structure, 280  
   substrate, 262, 263  
   surface, 256, 257  
   vaporisation, 303  
 Green function, 341  
 Green tensor, 150, 153, 154  
 group theory, 681, 701  
 growth modes, 48–52  
 guanine, 359  
 guanosine, 359  
 guest/host complex, 358, 359  
 gyromagnetic factor, 541  
 hard disk, 110, 260, 503, 547, 584–593, 616, 781  
   antiferromagnetically coupling media, 588–590  
   capacity, 616  
   discrete media, 592–593  
   magnetic, 585–588  
   perpendicularly magnetised media, 590–591  
   read head, 553  
   recording density, 616  
 hard-sphere potential, 751  
 harmonic generation, 684, 691, 702, 706, 734–736  
   second, 679, 714, 734–735  
   third, 679, 705, 736  
 Hartree–Fock theory, 429  
 hcp lattice, 183  
 Hebb rule, 792, 794  
 helicate, 355  
 helium, 246  
   ion, 30  
 Hellmann–Feynman theorem, 772  
 Helmholtz  
   energy, 425  
   equation, 138, 139  
 heteroatomic nanotube, 302  
 heteroepitaxy, 53  
 heterojunction, 336, 339, 408, 409  
 heterostructure, 162, 409  
   double, 410  
   III–V, 659  
   semiconductor, 764  
   Si/SiGe, 409–411  
   SiGe/Si/SiGe, 409  
 high-k insulator, 404  
 Hohenberg–Kohn theorem, 767–768  
 hole, 758, 761, 762  
   mobility, 386, 409, 410  
 holon, 343  
 HOMO, 343, 457, 458, 473–475, 769  
   delocalised, 472  
 HOPG, 256, 262, 263, 333  
   substrate, 257  
 Hubbard model, 461

- Hückel model, 196
- Hund rules, 507
- hybridisation, 190–191, 200, 261, 267, 270, 271, 280, 457, 528, 562
- hydrogen bond, 353, 354, 359–362, 367
- hydrophobic interaction, 354, 363–365
- hyper-Rayleigh scattering, 704
- hysteresis, 441, 442, 545, 546, 550, 551
  
- icosahedron, 184, 185, 187–189
  - gold, 190
- immunoglobulin, 689
- impurity, 435, 436, 543, 562, 652
- InAs quantum dots, 58, 59, 79, 80, 636–641, 643, 644
  - laser, 638
- inclusion complex, 364, 365
- indium, 375, 624, 640
- inelastic electron tunneling spectroscopy, 85
- InP, 261, 408, 409, 638, 652, 659
  - membrane, 651, 653, 656
  - microlaser, 659
  - nanowire, 495, 496
- insulator–metal transition, 264, 269
- integrated circuit, 383, 386, 783
  - BiCMOS, 409
  - CMOS, *see* CMOS
  - design parameters, 383
  - design rule, IX
  - half pitch, 395
- integration density, 260, 263, 383, 410, 616
  - memory, 600, 606, 609, 611
  - molecular memory, 477
  - of Josephson junctions, 445
- interdigital electrode, 165
- interface dislocation, 60
- iodobenzene, 88
- ion
  - bombardment, 11, 54, 246, 263
  - thinning, 36, 37
- ion beam etching, 12
- ion–ion interaction, 772
- ion–matter interaction, 35
- iron, 261, 510, 515–517, 561
  - atoms, 86, 331
  - whisker, 538
- island, 419, 422, 531
  - formation, 49–51, 53, 334
  - growth, 49
  - hexagonal array, 57
  - nucleation, 51, 52, 56
  - periodic growth, 56
  - polarised, 427
  - ramified, 256
  - size distribution, 52
  - stressed, 761
- isolated pentagon rule, 283
- itinerant magnetism, 510–511, 530, 542
- ITRS roadmap, IX, X, 395–397, 405, 440, 607
  
- Jahn–Teller effect, 216–217
- jellium model, 208, 210, 211, 220, 221, 223, 225, 232–235
- Josephson
  - current, 442–444
  - junction, 440–445
- Joule effect, 62, 246, 595, 609
  
- Kelvin force microscopy, 110
- Kerr effect, 540
- kinesin, 740, 741
- Kohn–Sham equations, 208, 768–770
- koiland, 365
- Krätschmer–Huffmann process, 284, 302
- Kretschmann prism, 624
- Kubo
  - criterion, 197–200, 202, 269–271
  - model, 202–205, 341
  
- Landau damping, 232, 233
- Landauer
  - four-wire formula, 431
  - linear response, 339
  - theory, 337, 430–432, 435
- Langevin process, 685, 701
- Langmuir film, 376
- Langmuir–Blodgett technique, 375, 472, 490
- laser ablation, 246, 494
- latching logic, 440, 442
- lateral force microscopy, 94
- latex film, 100, 114
- lattice mismatch, 49, 50, 53, 258, 259, 332, 529

- LCAO method, 292  
 LDOS, *see* local density of states  
 lead  
   nanotube, 114  
   nanowire, 333, 375  
 LECBD, 256–258, 262, 263  
 Legendre polynomial, 130  
 length scale, 671–673  
 Lennard-Jones potential, 93, 189, 750, 751  
 life, 674  
 life sciences, 671, 674  
 lift-off, 15–16, 33, 158, 162, 168, 328  
 light storage, 648  
 light-emitting components, 639  
 light-emitting diode, 639, 659  
   GaN, 639  
   organic, 674  
 lipid bilayer, 720, 726  
 liposomes, 370  
 liquid crystal, 364, 365, 736  
   display, 370  
 liquid metal ion source, 35  
 liquid-drop model, 180–181, 193–196, 245, 267, 682  
   quantum, 207, 213  
 lithium  
   cluster, 231  
   dielectric function, 232  
 lithography, 3–37, 42, 123, 327, 383  
   additive transfer, 15–18  
   AFM, 20  
   contact, 21–22  
   dip-pen, 115, 116, 491  
   dose, 5, 32  
   DUV, *see* lithography, EUV  
   electron beam, 4, 8, 30–35, 158, 168, 327, 452, 453, 493, 592, 593, 630, 636, 640, 650, 669  
   electron projection, 28–29  
   emerging techniques, 157–174, 326, 327  
   EUV, 20, 28, 327, 627  
   far-field techniques, 39  
   FIB, 35–39  
   field stitching, 31  
   industrialisation, 34  
   ion projection, 29  
   mask, 9, 11, 12, 14, 19–21, 25–28, 30  
   nanoimprint, *see* nanoimprinting  
   near-field, 172–173, 328–332, 342  
   next generation, 28  
   parallel writing, 19  
   pixel, 30  
   proximity, 22  
   resolution, 19, 21, 22, 24–26, 33, 38, 39  
   sequential writing, 19, 30  
   soft, 20, 40, 169–172, 327–328  
   subtractive transfer, 9–14  
   transfer, 8  
   write speed, 19, 38  
   X-ray, 18, 327  
 LLG equation, 541, 549  
 local density approximation, 766, 769–770  
   time-dependent, 232, 233, 238  
 local density of states, 73, 81, 457  
 localised magnetism, 508–510  
 lock-and-key mechanism, 370, 669, 686  
 logic  
   circuit, 497  
   gate, 483, 485, 497, 596, 606, 777, 782, 783, 794  
   operator, 782  
 look-up table, 787–789  
 Lorentz–Lorenz model, 679  
 LTP, 406  
 LUMO, 343, 457, 458, 473–475, 769  
   delocalised, 472  
 Luttinger liquid, 343  
  
 macroscopic quantum tunnel effect, 548  
 macrospin, 540, 576  
   LLG equation, 541  
   precession, 541–544  
 Madelung  
   constant, 192  
   energy, 192  
 magic cluster, 211, 212, 220, 221, 232, 273  
 magnesium cluster, 203  
 magnetic  
   alloy, 260  
   anisotropy, 260–262, 515–518, 543–545, 616, 773  
   cluster, 260–261, 531  
   dipole anisotropy, 521

- dipole interaction, 520–522, 524, 589
- domain, 520, 523, 585
- domain wall, 520, 523–524
- dots, 163, 260, 592, 593
- film, 529, 572–576
- fingerprint, 436
- flux quantum, 417, 434, 441, 443
- hard disk, 585–588
- head-to-tail wall, 524
- induction, 504, 623
- interface anisotropy, 526–531
- liquid, 371
- molecule, 548
- moment, 504–508, 773
- multilayer, 532–540, 552–572
- nanostructure, 164, 261
- RAM, *see* MRAM
- recording, 531
- recording head, 109
- semiconductor, 558
- shape anisotropy, 515, 524, 617
- storage, 547
- susceptibility, 271–272
- tape, 109
- thin film, 520, 522, 523, 595, 617
- tunnel junction, 557, 560, 568–572, 607
- ultrathin film, 523–524
- magnetic force microscope, 109, 163
- magnetisation, 512, 514–516, 520
  - dynamics, 540–551
  - easy axis, 260, 515, 516, 520, 523, 528, 544, 548, 549
  - easy direction, 516
  - hysteresis, 545, 546, 551
  - LLG equation, 541, 549
  - perpendicular, 590
  - precession, 541–544, 576–578
  - reversal, 544–551, 554, 576–578, 587
  - saturation, 588
  - spin transfer mechanism, 549–551
  - vortex, 521, 522
- magneto-optical
  - disk, 109
  - response, 123, 540
- magnetoconductance, 434–436
  - gold loop, 438
- magnetocrystalline anisotropy, 260, 507, 509, 510, 515–517, 521, 616
  - interface, 526–528, 574
  - uniaxial, 515, 520, 521, 576
- magnetoelastic anisotropy, 517
  - interface, 529–530
- magneto-electronic device, 504
- magneto-resistance, 550, 616
  - curve, 554, 556, 557
  - giant, 503, 533, 552–556, 560, 562–568, 586, 615
  - tunnel, 556–559, 568–572
- magneto-resistive device, 551
- magneto-resistive read head, 585, 616
  - giant, 585–586, 616
- magnetostatics, 504–505
- magnetostriction, 518–519, 543
- magnon, 543
- manipulation
  - of adatoms, 85
  - of atoms, 41, 86, 87, 331, 369
  - of molecules, 41, 87, 115, 369
- mass memory, 583–595, 781
  - local probe techniques, 584, 593–595
- matrix memory, 584, 595–613
  - access transistor, 604–608, 611
  - addressing, 598, 600
  - atomic scale, 613
  - bit line, 596, 601, 602
  - detection transistor, 611, 612
  - floating gate, 601, 602, 604, 612
  - gain cell, 611–613
  - NAND architecture, 602, 603
  - nanoscale, 599–606
  - noise, 599
  - NOR architecture, 602, 603
  - redundancy, 599
  - repair, 599
  - word line, 596, 602
- Maxwell's equations, 647
- Maxwell–Boltzmann distribution, 685
- Maxwell–Garnett model, 264
- MBE, *see* epitaxy, molecular beam
- McCumber parameter, 444
- mean field approximation, 759
- membrane, 369
  - cell, 370, 686, 720, 726–731
  - lipid, 369
- memory, 583–618
  - 64-bit, 497

- access time, 584, 600, 604, 606, 615, 780
- cache, 781
- cell, 22, 314, 344, 375, 417, 476, 478, 595, 596, 777, 783
- central, 781
- computer, 780–781
- flash, 584, 601, 603, 604
- hierarchical structure, 780, 781
- mass, *see* mass memory
- matrix, *see* matrix memory
- microprobe, 584, 594
- molecular, 163, 476, 477
- non-volatile, 477, 553, 598, 599, 601–604, 606
- NOVORAM, 606
- on-board, 599
- plane, 596, 598, 602, 605
- PLED, 612
- RAM, 598, 604–606
- register, 780, 781
- ROM, 596, 598, 601
- single-electron, 610–611
- stability, 584
- volatile, 604–605
- mercury, 199, 200, 269, 270
- mesophase, 370
  - lyotropic, 371
- mesoscale, 671
- mesoscopic
  - device, 417
  - system, 335
- metal–insulator transition, 198, 199
- metal–ligand bond, 354, 361
- metallic
  - binding, 184–189
  - carbon nanotube, 113, 297–299, 312, 487–489
  - cluster, 184, 196–200, 209, 213, 216, 225, 227, 232, 234, 253–254, 264
  - magnetic multilayer, 552–572
  - multilayer, 532
  - nanoparticle, 372, 449, 451, 452, 689, 698
- MFM, *see* magnetic force microscope
- micelle, 369, 370
  - reverse, 250
- micro–nano interconnects, 786
- micro-optics, 173
- micro-phase separation, 371, 374
- micro-squid, 16
- microcanonical ensemble, 754
- microcavity, 642, 643, 645–662
- microcontact printing, 40, 170, 171, 491
- microdisk, 645
- microelectronics, IX, 162, 261, 383–413, 783, 784
  - design, 615
  - memory requirements, 600
- microfluidic
  - channel, 165, 168, 495, 496
  - chip, 165
  - device, 170, 172, 627, 662
  - network, 171, 172
- microlaser, 646, 659
- micromagnetism, 173
- micropillar, 79, 643–645
  - GaAs/AlAs, 644
- microprocessor, 22
  - characteristics, 785
- micropump, 172
- microreactor, 370
- microscale, 671
- microsegregation, 371
- microtubule, 740, 741
- microvalve, 172
- Mie
  - classical theory, 226
  - resonance, 226–228, 233, 236–239, 253, 771
  - theory, 241, 265
- Miller indices, 42, 45, 182, 187
- Millipede, 173, 594
- miniaturisation, IX, 314, 599, 600, 619, 646
- miscut angle, 46
- misorientation angle, 59
- MOCVD, 55
- molecular
  - abacus, 87
  - amplifier, 474, 477
  - assembly, 490–492
  - bistable, 476, 478, 613
  - circuit, 490–498
  - components, 450–479, 497
  - conductor, 457–469
  - conformation, 470
  - diode, 449, 471–474

- dipole, 678
- electronics, 447–499, 795
- energy levels, 455
- engineering, 698–707
- machine, 366–368
- magnet, 371
- memory, 163, 476, 477
- motor, 738, 740, 741
- nanocage, 361
- network, 362
- orbitals, 457–463
- orientation, 676
- photonics, 667
- recognition, 251, 360, 362, 364, 370, 688
- rectifier, 449
- rotor, 366, 367
- shuttle, 366, 367
- switch, 373
- tectonics, 362, 363
- transistor, 477–479
- triode, 449
- tweezer, 358, 359
- wire, 78, 469–471
- molecular dynamics simulation, 202, 255, 258, 305, 341, 751–755
  - ab initio, 773
  - integration algorithm, 753–754
  - thermostat, 754–755
  - velocity renormalisation, 755
- molecule–metal coupling, 454–456, 459
- molybdenum cluster, 255
- monolayer, 49
- Monte Carlo simulation, 31, 32, 202, 755–757
  - selective sampling, 756
- Moore’s law, IX, 383, 395, 784
- MOSFET, IX, 162, 383–386
  - bulk punch-through, 392–394, 404
  - buried doping, 394, 404
  - conduction channel, 385
  - doping profile, 383, 393, 394, 405
  - drain, 383
  - electrical parameters, 396
  - gate oxide layer, 383
  - halo, 389, 394, 396, 404
  - inversion layer, 384
  - normally-off, 383–386
  - ohmic regime, 385, 391
  - on SOI, 412
  - pinch-off voltage, 385
  - pocket, 389, 394, 396
  - quantum effects, 406, 407
  - retrograde doping, 388, 393, 394, 396, 397, 404
  - saturation current, 385
  - saturation regime, 385, 390
  - scaling, 392–400
  - short channel effects, 392–393, 404, 604
  - silicon, 386, 483, 484
  - source, 383
  - space charge region, 384, 392
  - superhalo, 405
  - surface punch-through, 393, 411
  - threshold voltage, 384, 393
- Mott transition, 264
- MRAM, 371, 503, 504, 553, 596, 606–609, 613, 796
  - bit line, 608
- multi-quantum well, 656, 659
- multiphoton microscopy, 731–737
- multitwinned particle structure, 188, 191
- multiwalled nanotubes, *see* carbon nanotube, MWNT
- nano-antenna, 127
- nano-operator, 785
- nano-optics, 137, 144, 164
- nanosarray, 375
- nanobiophotonics, 667, 670, 686
- nanocage, 361
- nanochain, 375
- nanochannel, 165
- nanocluster, 186
- nanocomponents, 778
- nanocomputing, 776–798
- nanosconnection, 785
- nanocrater defect, 263
- nanodetector, 136
- nanoelectronics, 261, 383, 416–444
  - hybrid, 477, 497
  - superconducting, 440–444
- nanoembossing, 167–169, 328
- nanogap, 451, 456, 478, 479
- nanoimprinting, 20, 40, 158–166, 327, 328, 477, 593

- alignment, 162, 171
- UV, 166
- nanoindentation, 112
- nanolaboratory, 660, 662
- nanolithography, 4, 38, 39, 136, 157–174, 329
- nanomagnet, 371, 551
- nanomagnetism, 163–164, 504–551
  - fundamental lengths, 525
  - novel effects, 525–540
  - numerical simulation, 766
- nanomemory, 785
- nanoMOS devices, 400–412
- nanomoulding, 327, 328
- nanoparticle
  - assembly, 690
  - core–shell, 696
  - semiconductor, 697
- nanophotonics, 665–745
- nanopillar, 165, 551, 555, 556
- nanopore, 335, 451, 453, 469, 471
- nanoscale, 671
- nanoscale light source, 136
- nanosensor, 669, 700
- nanostencil, 115
- nanowire, 45, 324–344, 449, 553, 555, 631, 796
  - as electrical contact, 325
  - array, 375
  - as building block, 325
  - assembly, 493
  - bundle, 785
  - electrical conductance, 335–343
  - electrical contacts, 336–342
  - encapsulated, 313
  - fabrication, 326–327
  - InAs/InP, 335
  - InP, 495, 496
  - molecular, 342, 343
  - multilayer, 334
  - self-assembly, 332–334
  - semiconductor, 334
  - silicon, 115, 116, 495
  - ZnO, 336
- near-field microscopy, 41, 121–155, 312, 327, 328, 330, 332, 336, 627, 633
  - carbon nanotube tip, 317
- near-field optical microscope, 122, 123
  - apertureless, 123, 126–133
  - tip, 129–131, 135
- near-field optics, 121–155
- Néel pair model, 518–519, 528, 574
- Néel temperature, 508
- Néel wall, 524
- Néel–Brown model, 547
- negative refraction, 633
- Nernst–Einstein relation, 52
- neural networks, 791–794, 797
  - adaptive synapse, 794
  - formal neuron, 791
  - learning dynamics, 792
  - relaxation dynamics, 792
  - synapse, 791
- neuron, 686
- nickel, 113, 247, 261, 510, 515–517, 529, 539, 561
  - columns, 163
  - deposition, 16, 17
  - dots, 316
  - pillars, 593
- niobium
  - junction, 442, 445
  - layer, 553
- nitride, 605
- nitrogen film, 57
- NMOS, 385–387, 389–391, 393, 394, 410
- NMR, *see* nuclear magnetic resonance
- noble metal, 247, 537, 538
  - cluster, 237, 238
- non-covalent force, 354
- non-radiative
  - coupling, 147, 155
  - recombination, 639
  - transition, 681
- nonlinear optical effects, 660
- normal hydrogen electrode, 247
- NOVORAM, 606
- nuclear magnetic resonance, 109, 773
- nucleotide, 688
- nucleus, 207
- numerical aperture, 24, 25
- numerical simulation, 341, 749–774
  - accuracy, 773
  - computation time, 773
  - conjugate gradients, 752
  - effective mass method, 760–762
  - electron, 757–773
  - interatomic potential, 750–752



- nanomagnetism, 766
- Newton–Raphson method, 752
- potential energy surface, 752–753
- pseudopotential method, 760, 762–764, 766
- reliability, 773
- semi-empirical methods, 759–766
- steepest descent, 752
  
- octahedron, 184–186, 266
  - truncated, 184, 185
- odd–even effect, 197, 205
- off-axis illumination, 25
- oligomer
  - phenyl, 469
  - $\pi$ -conjugated, 469, 470
  - thiophene, 469
- oligophenylene
  - ethylene, 469, 470
  - vinylene, 469
- opal, 661, 662
- optical fibre, 123, 124, 627, 652, 667, 716, 736
  - metal-coated, 133–135
  - multimode, 627
  - sensor, 630
  - single-mode, 627
  - unconditionally secure link, 642
- optical multiplexing, 627
- optical tweezers, 737–740
- optoelectronic components, 22, 532, 641, 646
- optronics, 261, 619–662
  - optical addressing, 630, 631
  - S3PS, 641–644
  - subwavelength aperture, 627–629
- orange peel coupling, 575
- organic
  - dye, 688
  - nanostructure, 78, 79, 87
  - nanotube, 362
  - transistor, 163
- organometallic
  - complex, 450, 699
  - compound, 55
- organomineral, 699
- oxynitrides, 404, 605
  
- palladium cluster, 453
  
- parallel process, 19, 117, 161, 173
- paramagnetism, 587
- parameter mismatch, 50, 334
- parity, 681
- passivation layer, 14
- Pauli exclusion principle, 273, 429, 507, 758
- PBG, *see* photonic band gap
- PCRAM, 584, 596, 606–609, 613
- PDMS, 169–172, 327, 328
  - stamp, 169–171, 496
- Peierls transition, 343
- pentacene, 719
- perceptron, 792
- percolation threshold, 134, 258
- perturbation theory, 680, 770
- pH measurement, 707
- phase-shift mask, 25, 26, 30
- phonon, 437, 543, 755
- phospholipid, 370, 697
- phosphorescence, 693, 694
- photobleaching, 692, 709, 715, 716, 743
- photodepletion, 229
- photoemission spectroscopy, 458
- photoevaporation spectroscopy, 229–231
- photolithography, 20–28, 327
  - projection, 23–26
  - X-ray, 26–27
- photoluminescence, 261, 262, 640, 641, 643, 644, 659
- photomultiplier, 709
- photon
  - antibunching, 719
  - bunching, 719
- photon scanning tunneling microscope, 123–126
- photon sieve, 628–629
- photonic band gap, 630, 646, 655, 657, 658
  - 1D, 648
  - 2D, 649–650
  - absolute, 650
- photonic crystal, 164, 645–662
  - 1D, 647–648
  - 2D, 648–650, 652–655
  - allowed bands, 646, 653–656
  - coupling, 652
  - defect engineering, 660
  - dispersion relation, 647, 648, 651

- group velocity, 648, 654, 661
- laser, 654
- leaky mode, 652
- light cone, 651
- light line, 651
- mirror, 657
- optical confinement, 652–653, 656, 660
- strong confinement, 653
- vertical confinement, 652, 653
- weak confinement, 653
- with defects, 656–661
- photopolymerisation, 706
- photoresist, *see* resist
- photothermal detection, 698
- $\pi$  interaction, 354, 365, 367
- picotechnology, 499
- plane wave expansion, 137–142, 647
- plasma, 620
  - dispersion relation, 622
  - frequency, 622
  - reaction, 246
  - wavelength, 622
- plasmon, 126, 226–228, 234, 238, 239, 253, 276, 619–633, 698, 773
  - ATR coupling, 626
  - biochemical sensor, 624–626
  - bulk mode, 621
  - chemical sensor, 624
  - detection, 631, 632
  - dispersion relations, 622–623
  - fibre-optic sensor, 627
  - frequency, 629, 630
  - guide, 126
  - in metal nanoparticle, 629–633
  - light coupling, 622–627
  - polariton, 620
  - propagation length, 624, 633
  - resonance, 624
  - wave guide, 631–633
- platinum, 45, 330, 528, 592
  - acetylacetonate, 593
  - nanoparticle array, 16
  - nanowire, 331
  - shell, 248
  - sputtering, 453
  - surface, 56, 57, 531
  - vicinal surface, 333
- plexiglass, 7
- plumbago, 280
- PMMA, 7, 31–33, 36, 158, 160, 169, 327, 371, 375, 699
  - bridge, 454
  - resist, 636
- PMOS, 385–387, 390–391, 410
  - threshold voltage, 386
- point spread function, 713
- polarisability, 680
- polarisation tensor, 678, 705
- polariton, 677
- polaron, 343
- polycarbonate, 158
- polydimethylsiloxane, *see* PDMS
- polyhedron, 266, 282
  - compactness, 184, 185
  - Euler theorem, 191
  - regular, 185
  - symmetry group, 185
- polymer, 674
  - contour length, 108
  - dextran, 108
  - elasticity, 109
  - glass transition, 158
  - luminescent, 720
  - melt, 371
  - molecular weight, 7, 159
  - moulding, 40, 158–160, 166–168
  - photosensitive, 706
  - resist, 4, 158
  - silicone oil, 169
  - under irradiation, 7
  - viscosity, 159
  - wetting properties, 375
- polymethylmethacrylate, *see* PMMA
- polymolecular assembly, 368–378
- polystyrene, 371, 375, 737, 738, 741
- porous matrix, 327, 335, 336
- porphyrin, 357–360
- post-exposure bake, 5
- potassium, 485
  - cluster, 231
  - dielectric function, 232
- prepatterned surface, 41–65, 332, 553, 555
- prestructured surface, *see* prepatterned surface
- programmable
  - architecture, 497, 602

- ROM, *see* EPROM, EEPROM
- switch, 497
- protein, 673, 689, 695, 738, 741
  - conformation, 725
  - fluorescing, 719
  - folding, 725–727
- proximity effect, 31, 32
- proximity optical correction, 25
- pseudomorphic growth, 50
- pseudopotential method, 762–764, 766
- PSTM, *see* photon scanning tunneling microscope
- Pt/Co dots, 163
- Purcell effect, 643–645, 660
- pyrazine, 358
- pyridine, 253, 357, 358
- PZT, 606
  
- quantum box, *see* quantum dot
- quantum cascade laser, 631
- quantum coherent conductor, 338
- quantum computer, 642, 798
- quantum confinement, 422, 532–540, 575, 635
- quantum corral, 86
- quantum cryptography, 79, 641, 642, 644
- quantum data processing, 627, 645
- quantum dot, 50, 58–61, 79, 465, 466, 498, 639, 652, 761, 791
  - absorption spectrum, 640
  - laser, 634, 636, 638
  - plane, 635–637
  - radiative cascade, 642
  - semiconductor, 634–646, 660
  - single, 640–646
  - spectroscopy, 82–84
  - transistor, 163
- quantum interference, 428–439, 532, 534
  - Bohm–Aharonov effect, 437–439
- quantum size effects, 533, 537–540, 575–576
- quantum well, 635, 637, 639, 640, 642, 652
  - components, 639
  - laser, 634, 635, 638, 639, 761
- quartz, 434
  - template, 167, 168
- qubit, 642
  
- radiative
  - cascade, 642
  - coupling, 147, 155
  - damping, 147–149
  - lifetime, 148, 151, 152, 154, 643
  - mode, 650, 651, 653, 654
  - recombination, 639, 643
- radiolytic effect, 33
- RAM, *see* memory, RAM
- Raman
  - microspectrometry, 630
  - scattering, 708, 709, 714, 736
  - spectroscopy, 310, 683, 708
  - spectrum, 773
- random walk, 433
- Rayleigh criterion, 24, 142, 619, 627, 713
- Rayleigh scattering, 708
- reactive ion etching, 13–14, 25, 36, 158, 162, 168, 389, 453, 636
  - etch rate, 13
- reciprocal lattice, 534, 536–538, 647, 648
- reconstructed surface, *see* surface reconstruction
- recording density, 260, 503, 547, 584, 585, 588, 594, 595, 616, 780
- reduction potential, 247
- refractive index, 123, 648
  - contrast, 652, 660, 661
  - high, 652
- reorientation transition, 528
- replication, 687
- resist, 3–9, 19, 593
  - contrast, 6, 20
  - contrast curve, 7
  - electrosensitive, 30
  - exposure, 4
  - glass transition temperature, 15
  - inorganic, 33, 37
  - negative, 6, 31
  - polymer, 158
  - positive, 6–8, 31
  - refractive index, 21
  - sensitivity, 5, 7
  - sol-gel, 166
  - spreading, 5
- Reststrahlen region, 621
- Rhodamine 6G, 253, 694, 723

- RKKY interaction, 540, 589, 590, 617  
 RNA, 353, 720, 742  
 ROM, *see* memory, ROM  
 rotaxane, 356, 366, 367, 476, 478, 497  
 RRK model, 245  
 RSFQ  
   components, 442–445  
   logic, 417, 440  
   logic gate, 443, 444  
 ruthenium, 450, 589  
 ruthenium–trisbipyridine complex, 699
- sapphire, 640  
 SBT, 606  
 SCALPEL, 29  
 scanning capacitance microscopy, 111  
 scanning electron microscope, 31, 79, 316  
 scanning near-field optical microscope, 173, 715  
   aperture, 133–136  
   apertureless, 131–133, 630  
 scanning projection printing, 23  
 scanning spreading resistance microscopy, 111  
 scanning tunneling microscope, 4, 41, 68–89, 123, 173, 309, 311, 329, 450, 667, 668  
   barrier height, 76–77, 82  
   contrast, 75–76  
   elastic tunnel current, 80–82, 374  
   for biology, 741  
   image potential, 77  
   inelastic tunnel current, 84  
   local chemistry, 87–88  
   manipulation, 85–87, 368  
   modelling, 773  
   pulling mode, 85  
   pushing mode, 85  
   resolution, 74–75  
   setup, 71  
   single-molecule observation, 451–452  
   sliding mode, 86  
   spectroscopy, 80–85, 766  
   tip apex, 72, 75, 76  
   tip preparation, 71–72  
   tip–sample interaction, 85–88  
   topography, 70, 374  
 scattering matrix, 337, 431, 650  
 scattering sphere model, 127–128  
 Schottky  
   barrier, 314, 480–483  
   diode, 314  
   gate electrode, 409  
 screw dislocation, 60  
 selection rules, 681  
 selenium, 470  
 self-assembled monolayer, 4, 170, 491–492, 495  
 self-assembly, 342, 498, 661, 669, 688  
   by charge transfer, 365  
   by hydrophobic interaction, 363–365  
   FePt nanoparticles, 592, 593  
   hydrogen bonding, 359–362  
   in bulk, 369–371  
   of amphiphiles, 370, 371  
   of carbon nanotubes, 300  
   of nanowire, 332–334  
   on surface, 372–378  
   techniques, 327  
   template effect, 354–359  
 self-consistent calculation, 208, 232, 238, 239, 462, 472, 766, 770–771  
 self-organisation, *see* self-assembly  
 self-organised growth, 79, 636  
 semi-empirical methods, 457  
 semiconductor, 50  
   band gap, 80–82, 761  
   carbon nanotube, 298–300, 314  
   cavity, 645  
   doping, 109  
   Fermi wavelength, 418  
   ferromagnetic, 559  
   growth, 334  
   heterostructure, 764  
   III–V, 261, 639, 652  
   laser, 634–640  
   membrane, 652  
   multilayer, 532  
   nanoparticle, 449  
   nanostructure, 261  
   quantum dot, 634–646, 660  
   surface, 51  
   valence band, 765  
 SERS, *see* surface enhanced Raman spectroscopy  
 shaped-beam machine, 34  
 sharp-point effect, 129

- SHG microscopy, 734, 735  
short channel effects, *see* MOSFET,  
    short channel effects  
shot noise, 132  
Sigmund formula, 11, 12  
silane, 372  
silanisation, 165, 687, 697  
silica, 169, 479  
    insulator, 479  
    metallised grating, 627  
    sphere, 661  
    substrate, 496  
    surface, 372  
silicon, 11, 25, 639, 652  
    atomic orbitals, 764  
    band gap, 261  
    band structure, 762, 765  
    bead, 97, 107, 108  
    cluster, 261  
    dangling bond, 87  
    dots, 315  
    etched, 10, 14  
    hydrogenated surface, 4, 330  
    islands, 60  
    lines, 162  
    matrix, 163  
    monocrystalline, 4  
    MOSFET, 385  
    mould, 167, 169  
    nanocrystal, 765, 774  
    nanowire, 115, 116, 330, 495  
    oxidation, 55, 115, 388  
    oxide, 165, 170, 385  
    polycrystalline, 12, 169, 389, 402, 407  
    prepatterned surface, 64  
    probe, 594  
    pyramid, 38  
    substrate, 31, 32, 163, 169, 335, 495  
    surface, 77–79, 88, 107  
    thin film, 60  
    transistor, 163  
    vicinal surface, 61–63, 333  
    wafer, 28, 59, 60, 166, 168, 616, 627  
silicon-on-insulator, *see* SOI  
silver, 237, 330, 538, 624, 697  
    cluster, 205, 238–240, 249, 265  
    dielectric function, 238  
    film, 57  
    islands, 56  
    nanoparticle, 154, 689  
    nanowire, 333, 375  
    stripes, 633  
    sulfide, 250  
    surface, 372  
silylation, 25  
SIMS, 406  
single molecule, 136  
    adhesion, 108  
    component, 449  
    conductance, 452, 453, 457, 469, 470  
    elasticity, 108  
    electrical contact, 450–456  
    electronic device, 173  
    emissions, 144  
    fluorescence, 137  
    fluorescence detection, 707–731  
    level broadening, 457, 458, 463, 470  
    low temperature spectroscopy,  
        718–719  
    observation, 630, 675  
    spectroscopy, 709, 725  
    strongly coupled, 458–463  
    transistor, 451  
    transport properties, 456–469  
    weakly coupled, 463–469  
single-electron memory, 610–611  
single-electron transistor, 16, 417,  
    427–428, 449, 611, 612, 617, 794  
    CNT, 486–489  
    conductance, 487  
    stability diagram, 428, 487, 488  
single-photon source, 79, 641–644  
single-walled nanotubes, *see* carbon  
    nanotube, SWNT  
SNOM, *see* scanning near-field optical  
    microscope  
sodium, 237  
    borohydride, 247  
    cluster, 209, 212, 215, 219–221, 233,  
        234, 771  
    dielectric function, 233  
    dodecylsulfate, 248  
    magic cluster, 211  
soft bake, 5, 158  
soft matter, 169  
SOI, 412, 652, 659  
    layer, 162  
    technology, 411

- transistor, 609, 612
- solar energy, 303, 304, 306, 360
- soliton, 343
- spaser, 633
- spectral hole burning, 675
- spectral matching, 643
- spherical harmonic, 130
- spillout, *see* electron spillout
- spin, 506
- spin electronics, 552–572
  - CIP geometry, 553–556, 563–565
  - CPP geometry, 553–556, 563–565, 567
  - Mott model, 560–562
  - spin accumulation, 554, 567–568
  - two-current model, 559–562
- spin glass, 262
- spin transition complex, 377
- spin valve, 163, 553, 554, 586
- spin–orbit interaction, 506–507, 510, 528, 766
- spin–spin interaction, 766
- spinon, 343
- spintronics, *see* spin electronics
- spontaneous emission, 150, 151
- sputtering, 11, 15, 54, 55
  - etch rate, 12
  - platinum, 453
  - yield, 11–13
- squeezed field, 632
- SRAM, 483, 584, 600, 605, 606, 781
  - bit line, 605
  - switch, 596
- stacking fault, 44, 45, 56–58
- stamp, 328
- Stark effect, 719
- statistical physics, 755, 756
- step bunching, 61, 62, 64
- step edge, 57, 58, 78, 333
  - defect, 56
  - in magnetic film, 574–575
- step-and-flash, 166
- step-and-repeat projection printing, 23, 28, 29
- step-and-scan projection printing, 23
- STM, *see* scanning tunneling microscope
- STM-assisted CVD, 329
- stochastic matrix theory, 436
- Stokes shift, 694
- Stone–Wales transformation, 755, 756
- Stoner model, 510, 559
- Stoner–Wohlfarth astroid, 545, 546
- stopping power, 11
- Stranski–Krastanov growth, 49–51, 58, 334
- streptavidin–biotin pair, 686, 688, 691
- subnanoscale, 671
- supercapacitor, 317
- superconducting
  - logic components, 440–442
  - materials, 417
  - MWNT, 312
  - nanoelectronics, 440–444
  - niobium layer, 553
  - transition temperature, 773
- superconductor, 109
- superlattice, 371
- superparamagnetism, 260, 586, 588, 591, 595, 617
- supershell, *see* electron supershell
  - structure
- supersonic beam, 242
- supramolecular grid, 377
- supramolecule, 353–368, 674, 686
- surface
  - charge, 622
  - chemical potential, 52, 53, 58
  - curvature, 53, 58, 60
  - dehydrogenation, 87, 88
  - diffusion, 53, 57
  - diffusion coefficient, 52
  - dislocation, 44, 45, 50, 56
  - faceted, 45, 48
  - free energy, 46, 48, 49, 52, 53, 180, 258
  - functionalisation, 494, 495
  - functionalised, 669
  - graphite, 256, 257
  - plasmon, 126, 227, 228, 234, 238, 239, 619–633
  - reconstruction, 43, 44, 332, 750
  - relaxation, 42, 59
  - self-organised, 47–48, 56–57
  - stepped, 45
  - stress, 46, 47, 53, 57, 60
  - tension, 180
  - topography, 70, 105, 132, 165
  - vicinal, *see* vicinal surface

- surface enhanced Raman scattering, 630
- surface enhanced Raman spectroscopy, 253, 633
- surface-emitting laser, 645
- surfactant, 248, 249, 371, 372
- symbiotic computer, 798
- synchrotron radiation, 27
- systems on-chip, 614
  
- tailor-made molecule, 698
- TBSMA, 269
- TD LDA, 232, 233, 238
- technological node, 395
- tecton, 362, 363
- template effect, *see* self-assembly, template effect
- Teramac, 789
- Tersoff–Hamann approximation, 773
- Tersoff–Hamann theory, 73, 75
- terthiophene, 469, 470
  - dithiol, 470
- tetrahedron, 185
- THG microscopy, 736
- thienylenevinylene, 78
- thioalkane, 254
- thiol, 115, 116, 170, 173, 328, 330, 372, 374, 469, 688
  - nanowire, 330
- Thomas–Fermi approximation, 767
- Thomas–Reiche–Kuhn sum rule, 235
- THz source, 631
- tight-binding approximation, 184, 269, 292, 293, 300, 341, 760, 762, 764–766
  - self-consistent, 472
- time reversal symmetry, 434, 505, 515
- time-of-flight spectrometer, 214, 273
- tin, 375
- top-down approach, 41, 326, 327, 349, 688
- top-gate configuration, 480
- transferability criterion, 763
- transistor, 110, 157, 375, 761, 777, 783
  - as switch, 783
  - bipolar heterojunction, 409
  - buried oxide, 411, 412
  - CNTFET, 479–486, 494, 785
  - CNTSET, 486–489, 497
  - FET, 314, 611, 612, 615
  - FinFET, 412, 413
  - gate length, IX, 383
  - geometry, 392
  - HEMT, 408–409
  - interconnects, 398–400, 407
  - leakage current, 388, 390, 392, 394, 403, 404, 411
  - miniaturisation, 784
  - MOSFET, *see* MOSFET
  - multiple gate, 412
  - new architectures, 408–412
  - organic, 163
  - quantum dot, 163
  - SBFET, 480
  - SESO, 612
  - single-electron, 16, 427–428, 449, 486–489, 611, 612, 617, 794
  - single-molecule, 451, 477–479
  - SOI, 609
  - spinFET, 558, 560
  - threshold voltage, 384
  - ultra-high speed, 18
- transition metal, 261, 510
  - cluster, 199, 766
  - ferromagnetic, 517, 559
- transmission electron microscope, 36, 308, 310, 317
- triboelectricity, 109, 111
- tribology, 112, 591
- trimethylammonium bromide, 248
- tritopic ligand, 355
- truncated octahedron, 186
- tungsten
  - AFM tip, 103
  - STM tip, 71
- tunnel
  - current, 70, 72–79, 374, 403, 773
  - double junction, 425–427, 455
  - effect, 69, 72, 124, 410, 412, 420–422, 425, 463, 467, 540, 602
  - junction, 421, 422, 456, 557, 560, 568–572, 607, 612
  - magnetoresistance, 556–559, 568–572
  - resistance, 424, 425
- Turing machine, 779, 780
- twist angle, 60
- two-level model, 680, 686, 701
- two-photon acid photogenerator, 706

- two-photon fluorescence microscopy, 711, 732–734
- ultrasonic waves, 33
- uncertainty relation, 140, 141, 143, 425
- UV-NIL, *see* nanoimprinting, UV
- vacuum vapour deposition, 15, 53
- van der Waals force, 33, 93, 99, 109, 189, 246, 251
- Verlet algorithm, 753
- vertical-cavity semiconductor laser, 38
- vesicle, 370, 720, 736
  - bilayer, 370
  - giant, 735
- vicinal surface, 45, 46, 332, 333, 574, 576
  - gold, 48
  - growth on, 57–58
  - platinum, 531
  - silicon, 61–63
- virus, 369, 691, 730, 732
- VLS synthesis, 306, 307, 327
- Volmer–Weber growth, 49
- wave guide, 623, 627, 646, 651, 652, 656, 660–661
  - conical, 135, 136
  - metallic, 134, 135
  - surface plasmon, 631–633
- wave–particle duality, 761
- wet etching, 9–11, 330
  - undercut, 10
- wetting, 49, 375
- Wigner function, 677
- Wigner–Seitz
  - polyhedron, 183
  - radius, 180, 208, 238
  - unit cell, 183
- WKB approximation, 74
- Woods–Saxon potential, 207, 209, 211
- work function, 72, 196, 213, 267–268, 464, 483, 485
- Wulff polyhedron, 182–184, 189, 267
- YBaCuO technology, 445
- Zeeman
  - effect, 203
  - energy, 514
- zinc, 247