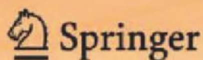
Nanoscience

Nanotechnologies and Nanophysics



Copyrighted material

Claire Dupas, PhD
Ecole Normale Supérieure de Cachan
Avenue du Président Wilson, 94235 Cachan Cédex, France
E-mail: 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

Marcel Lahmani, PhD
Club Nano-Micro-Technologie de Paris
Boulevard François Mitterrand, 91025 Évry Cédex, France
E-mail: 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

© 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

Copyrighted material

Part I

Tools for Nanoscience

Copyrighted material

1

Lithography and Etching Processes

D. Mailly 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.

Copyrighted material

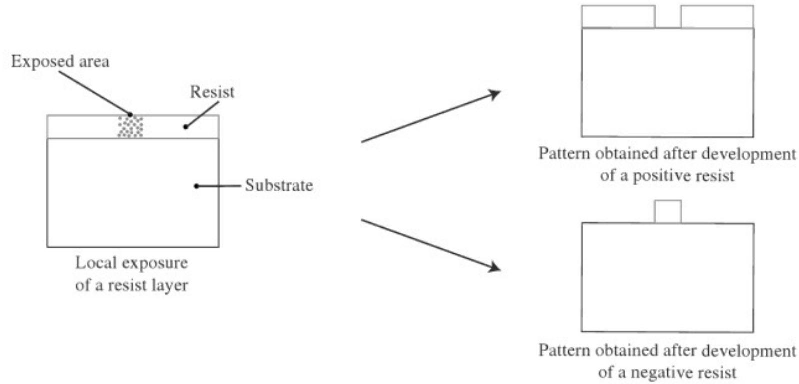


Fig. 1.1. Local exposure of resist and development

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 AlF_3, SiO_2 and 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

Copyrighted material

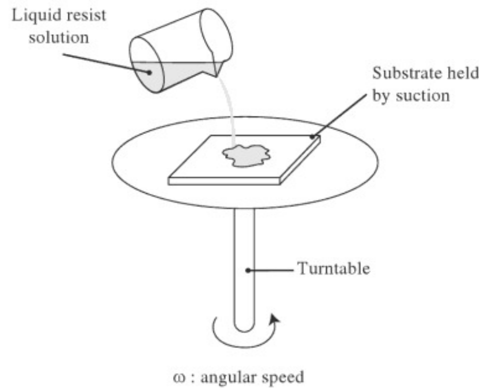


Fig. 1.2. Spreading the resist layer using a turntable

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°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°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

Copyrighted material

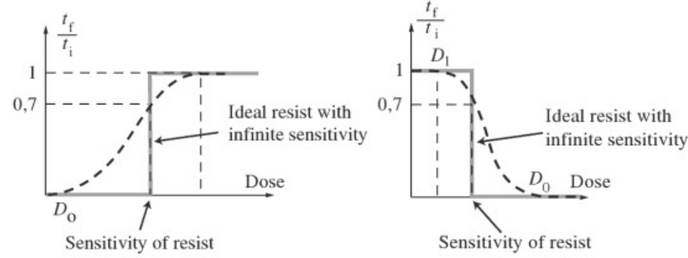


Fig. 1.3. *Left*: Negative resist. The resist hardens during exposure and the contrast curve increases. *Right*: Positive resist. The resist is softened during exposure and the contrast curve decreases

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^2. When the resist is irradiated by a photon beam, typical units are J/cm^2.

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.

Copyrighted material

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 γ 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, ε is the energy deposited per unit volume during exposure, ρ 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

Copyrighted material

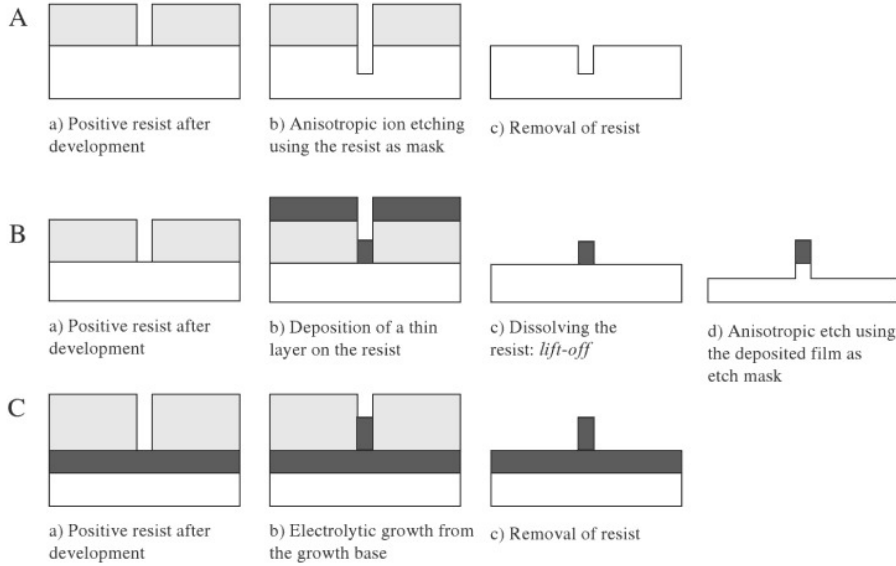


Fig. 1.4. Three examples of transfer strategies starting from a pattern obtained by lithography on a positive resist

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

$$V = V_0 + \beta / M_{\mathrm{f}}^{\alpha}$$

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 β and α 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,

Copyrighted material

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 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.

Copyrighted material

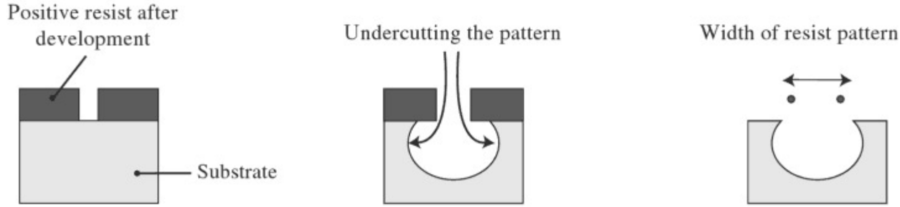


Fig. 1.5. Isotropic effect of a wet etch, e.g., when silicon is etched with a mixture of nitric acid, hydrofluoric acid and water

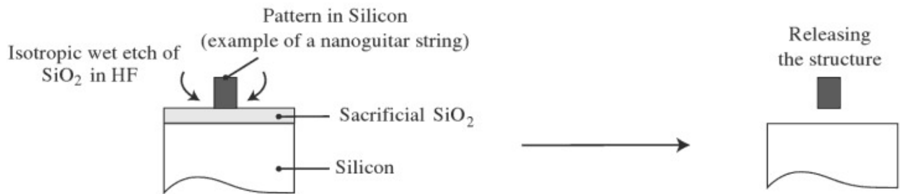


Fig. 1.6. Undercutting a nanostructure by wet-etching a sacrificial SiO_2 layer with a mixture of hydrofluoric acid and water. The endpoints of the freed structure shown here in cross-section rest on the substrate at anchoring points that have not been represented

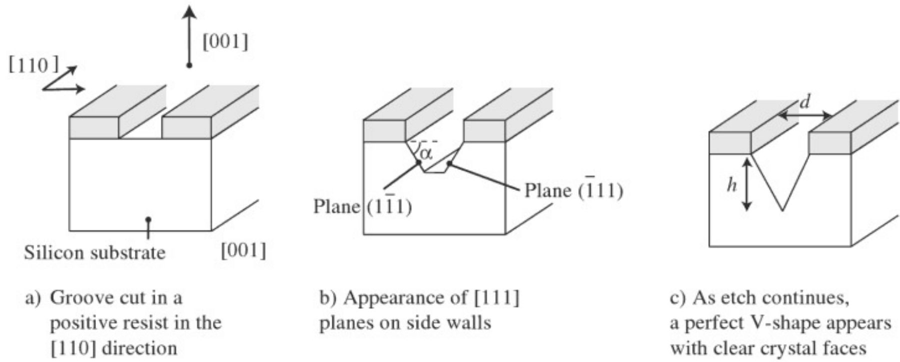


Fig. 1.7. Anisotropic wet etch of monocrystalline silicon [001] in an etchant of type KOH + water. The etch rate of this solution is negligible along the planes $\{111\}$. With this configuration, a V-shaped profile is etched out when the pattern edges are aligned with a surface of type [110], and undercutting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

Copyrighted material

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

$$S = \frac{3}{4}\frac{E_\mathrm{d}}{N\pi^2 CU}\,, \tag{1.1}$$

where $C = 1.81\,\mathrm{nm}^2$ is a constant, N is the atomic density of the material (in atoms/cm^3), 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.

Copyrighted material

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.[1] 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 $1\,\mathrm{mA/cm^2}$, 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

[1] 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.

Copyrighted material

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 anistropy.

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.

Copyrighted material

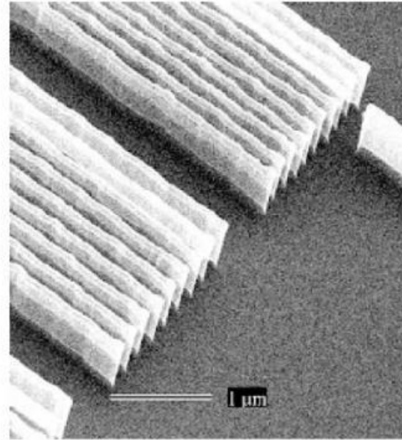


Fig. 1.8. Silicon walls etched by anisotropic RIE. Note the verticality of the sides. The walls have width 30 nm and height 600 nm, giving an aspect ratio of 1/20. The SEM (scanning electron microscope) image was taken at an angle to show that the very thin walls are transparent to the electron beam 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 a 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.

Copyrighted material

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.

Copyrighted material

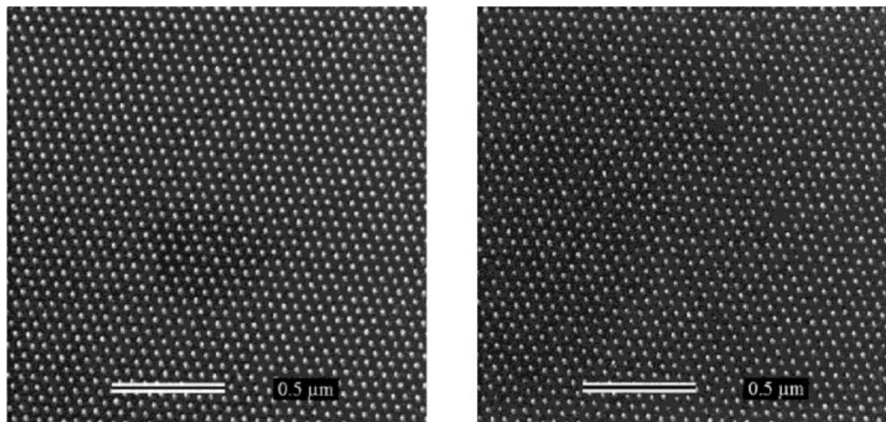


Fig. 1.9. Pt nanoparticle array fabricated by electron lithography and lift-off of a 6-nm Pt film. The Pt particles of diameter 10 nm (*left*) and $<$ 10 nm (*right*) are arranged in a lattice of period 50 nm. Photo F. Carcenac (LAAS/CNRS)

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).

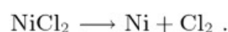
1.4.2 Electrolytic Growth

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, Ni^{2+} ions in the electrolyte are reduced ($Ni^{2+} + 2e^- \rightarrow Ni$) and deposited on the sample surface, whilst at the anode, Cl^- ions are oxidised ($2Cl^- \rightarrow Cl_2 + 2e^-$), and chlorine gas given off. The overall redox reaction is thus



$$NiCl_2 \longrightarrow Ni + Cl_2 \ .$$

Copyrighted material

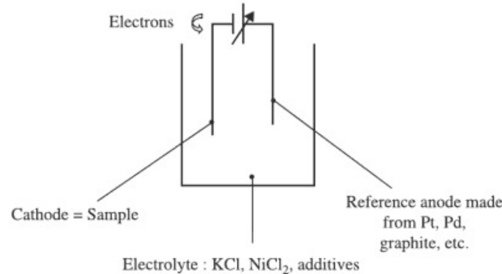


Fig. 1.10. Electrolytic cell for nickel deposition

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

$$m = \frac{ItM}{\mathrm{F}z} ,$$

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.

Copyrighted material

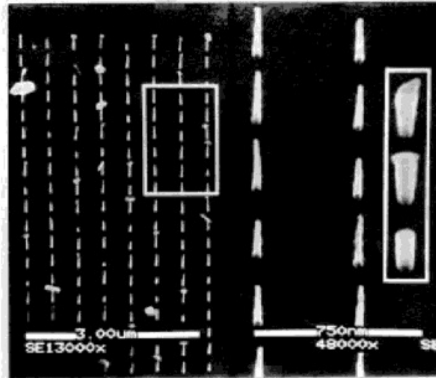


Fig. 1.11. Gold pillars obtained by electron lithography and electrolytic growth. The pillars have diameter 30 nm. In these tiny nanostructures, the growth rate from one object to another is not the same. Some pillars are only 50 nm high, whereas others exceed 300 nm. In the latter case, electrolytic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

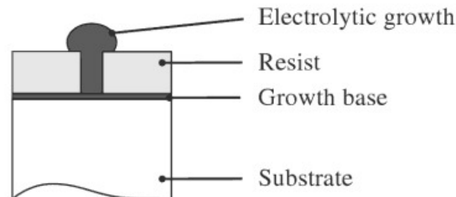


Fig. 1.12. Typical shape of a pattern feature obtained when electrodeposition has overflowed the resist mould. This mushroom shape has been deliberately used to make gates for ultra-high speed transistors

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.

Copyrighted material

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 ($> 50\,\mathrm{nm}$), limited by diffraction effects. On the other hand, sequential methods like electron beam lithography with very high resolution ($< 10\,\mathrm{nm}$) 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

Copyrighted material

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

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

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

Copyrighted material

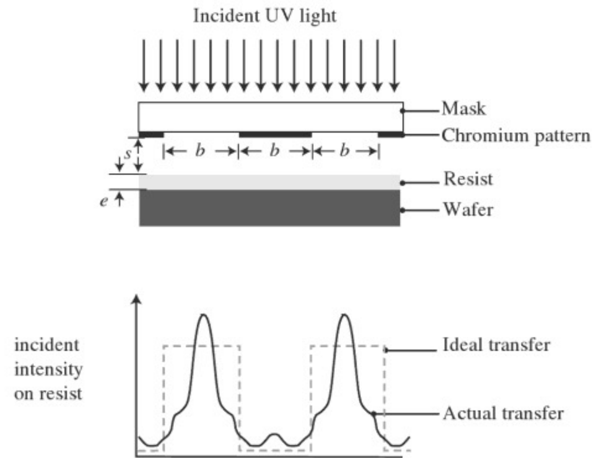


Fig. 1.13. Light intensity profile on a resist layer, obtained with a mask carrying a parallel array of equally spaced bands of width b. The *dotted lines* show ideal light transfer and the *continuous curve* is the true light profile on the resist layer

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 μm, 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 μm.

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

Copyrighted material

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 µm.

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 µm in a resist layer of thickness 0.5 µm. Using thinner resist layers and an F_2 excimer laser which delivers light at 157 nm, features of size 150 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 µm, the maximal resolution with a wavelength of 400 nm is of the order of 3 µm. 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 µm. 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.

Copyrighted material

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.

Copyrighted material

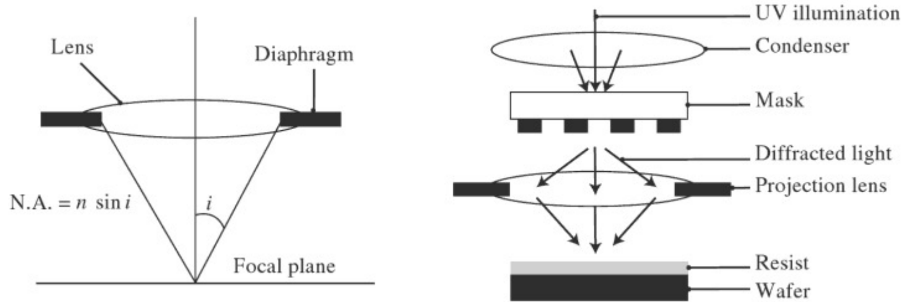


Fig. 1.14. Definition of the numerical aperture (N.A.) and optics with mask and wafer

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 λ 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

Copyrighted material

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

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

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 μm 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_2 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.

Copyrighted material

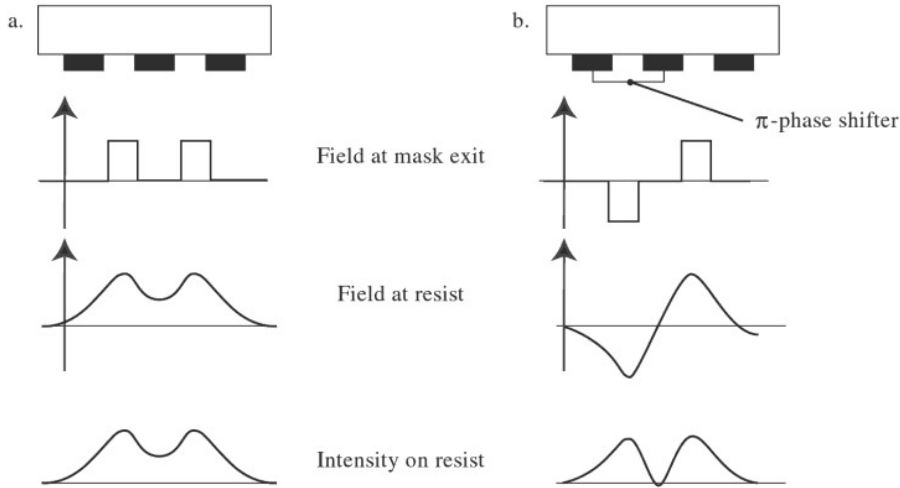


Fig. 1.15. Phase-shift mask. (**a**) Standard mask. Due to diffraction effects, there is a nonzero intensity at a position corresponding to an opaque part of the mask. (**b**) The π-phase shifter modifies the amplitude of the electric field and thus strengthens the shadow zones

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

Copyrighted material

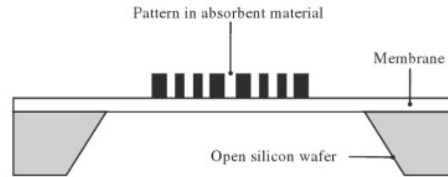


Fig. 1.16. Mask for X-ray lithography

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 μm. 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.

Copyrighted material

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

Copyrighted material

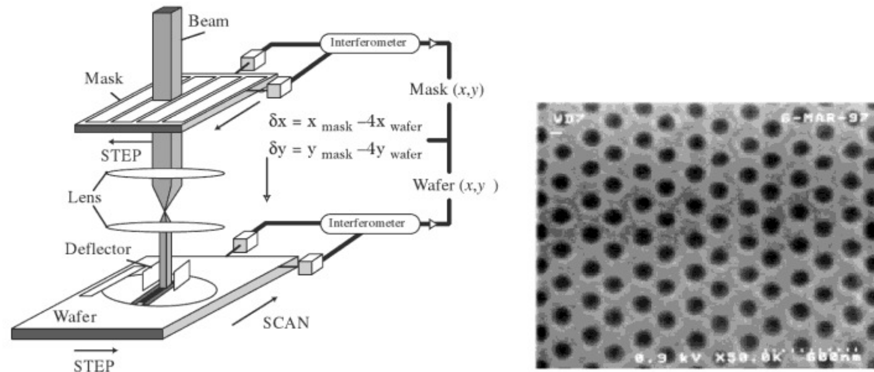


Fig. 1.17. *Left*: Schematic view of the SCALPEL electron projection setup with step-and-repeat. *Right*: Example of SCALPEL production. Holes of diameter 80 nm in a DUV resist of thickness 750 nm. Courtesy of L.R. Harriot, Bell Laboratories: `http://accelconf.web.cern.ch/AccelConf/p99/PAPERS/FRBL1.PDF`

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 $1\,\mu\mathrm{C/cm^2}$ compared with several tens of $\mu\mathrm{C/cm^2}$ for electrons.

Copyrighted material

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., $\lambda\,(\mathrm{nm}) = 1.22/\sqrt{E\,(\mathrm{eV})}$, 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 μm × 100 μm, 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

Copyrighted material

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-μm 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

Copyrighted material

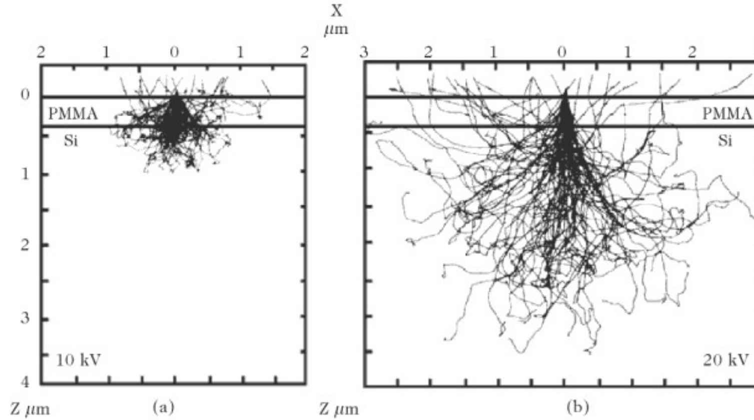


Fig. 1.18. Monte Carlo simulations of the trajectories of 10-keV and 20-keV electrons arriving at a point on a silicon substrate coated with 0.4 μm of PMMA [2]

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

Copyrighted material

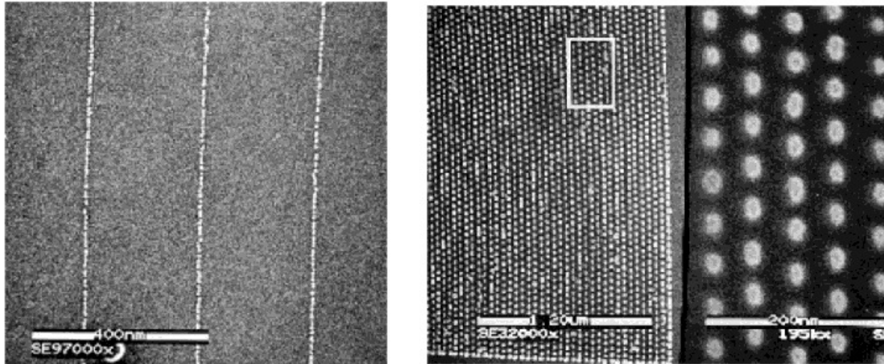


Fig. 1.19. Example of resolution limit on PMMA. *Left*: Isolated lines of width 7 nm obtained by gold lift-off. *Right*: SiO_2 dot array with period 40 nm obtained by lift-off followed by etching. Photos C. Vieu

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.

Copyrighted material

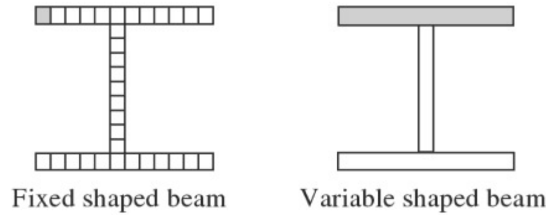


Fig. 1.20. Writing strategies for shaped-beam machines

The stoichiometry of certain oxides such as or 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 Al_2O_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

Copyrighted material

are fabricated with this type of equipment. Of course, the resolution limit is less good than with a Gaussian beam, being typically 0.20 μm, 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 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°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 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

Copyrighted material

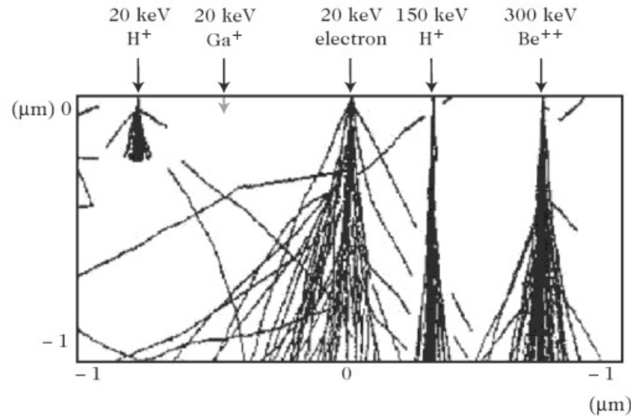


Fig. 1.21. Absorption of ion energy for ions with different masses and energies in a 1-μm PMMA layer and comparison with electrons at 20 keV

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

Copyrighted material

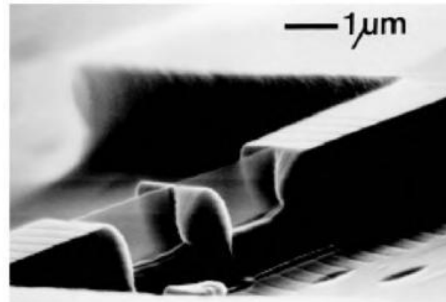


Fig. 1.22. Thinning of a GaAs mesa for TEM observation

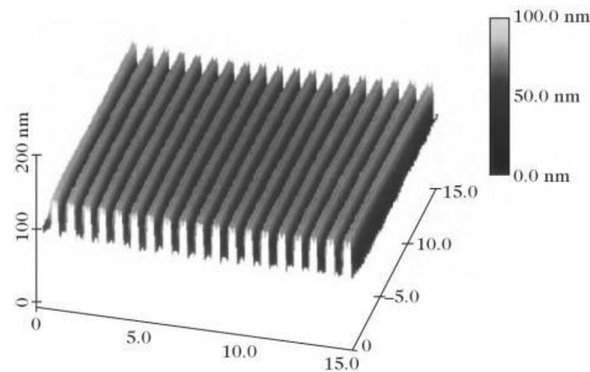


Fig. 1.23. AFM image of a series of lines obtained by FIB lithography on AlF_3. Note the good verticality of the lines

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, 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.

Copyrighted material

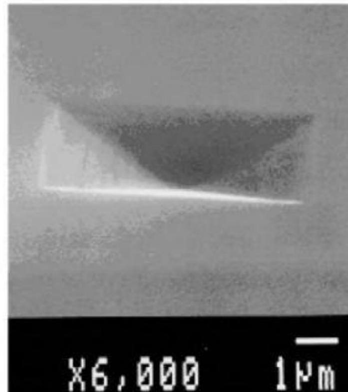


Fig. 1.24. Pyramid carved out of a a silicon substrate by modulating the dose during scanning by the FIB

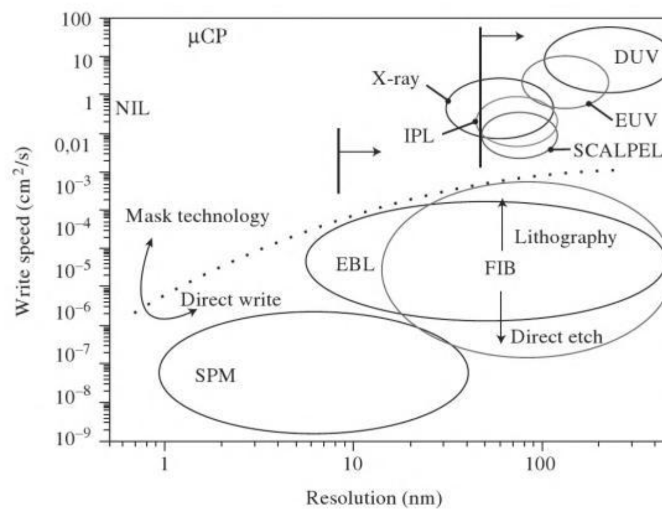


Fig. 1.25. Panorama of nanolithographic methods on a graph of write speed versus spatial resolution

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

Copyrighted material

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 nm × 70 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 μm	Laboratory and R&D	Economical
	Proximity	2 μm	Laboratory and R&D	Economical but low resolution
	Projection	80 nm	Industry	Spectacular progress
X-ray lithography		30 nm	Not used at the present time	Development suspended
EUV lithography		$<$ 50 nm	Industry	Could represent next generation in 2005
Electron lithography	Focussed beam	1 nm	Laboratory and R&D. Optical mask fabrication	Technique without mask. Best resolution
	Projection	50 nm	Demonstration	Many difficulties remain
Ion lithography	Focussed beam	8 nm	Demonstration	Better suited to etching than to lithography
	Projection	50 nm	Demonstration	Many difficulties remain. Still immature

Copyrighted material

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 $1\,\mathrm{cm}^2/\mathrm{s}$. It is immediately clear from this graph that there is a problem when we wish to combine such speeds with resolutions close to 10 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 μCP) 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)

Copyrighted material

2

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 μm 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

Copyrighted material

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

Copyrighted material

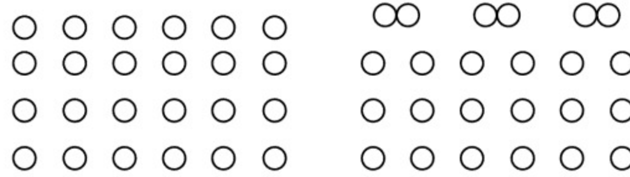


Fig. 2.1. *Left*: Cross-sectional view of normal relaxation. *Right*: Cross-sectional view of reconstruction. From [3]

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 1×1 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

Copyrighted material

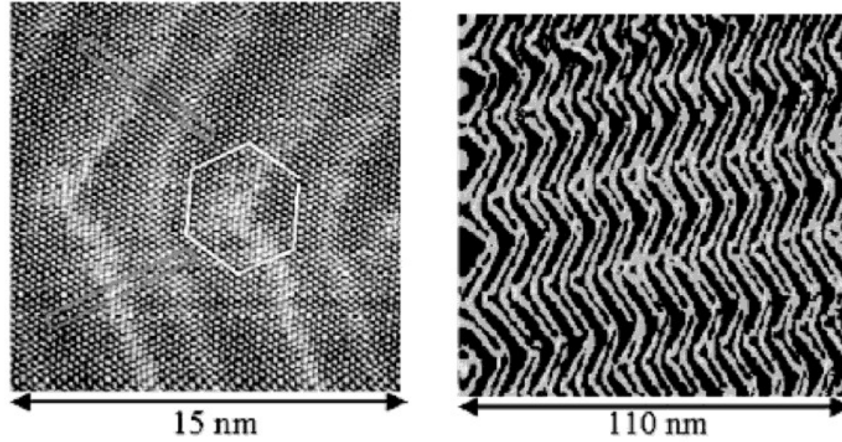


Fig. 2.2. *Left*: STM image of the gold surface Au(111) showing the $22 \times \sqrt{3}$ reconstruction. Two reconstructed domains are shown by *rectangles*. The *hexagon* is a Burger circuit showing the core of a surface dislocation at the join between reconstructed domains. Photo S. Rousset. *Right*: 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

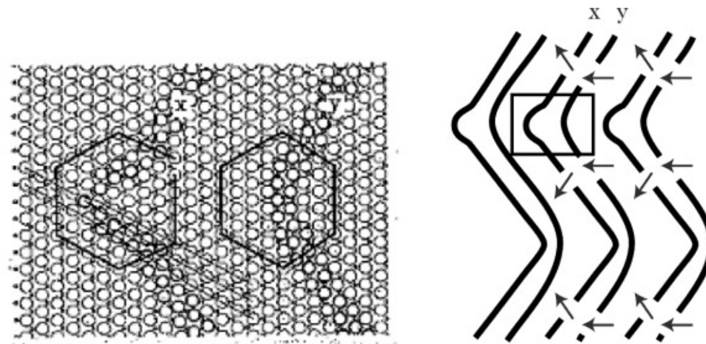


Fig. 2.3. *Left*: Atomic structure of a bend in the herringbone reconstruction, showing a surface dislocation characterised by its Burger circuit *on the left*, marked by x. From [4]. *Right*: Lines of stacking faults organised into zigzags. Bends x and y have different structures. The *rectangle* indicates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

Copyrighted material

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° 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° 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°. 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 θ (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

Copyrighted material

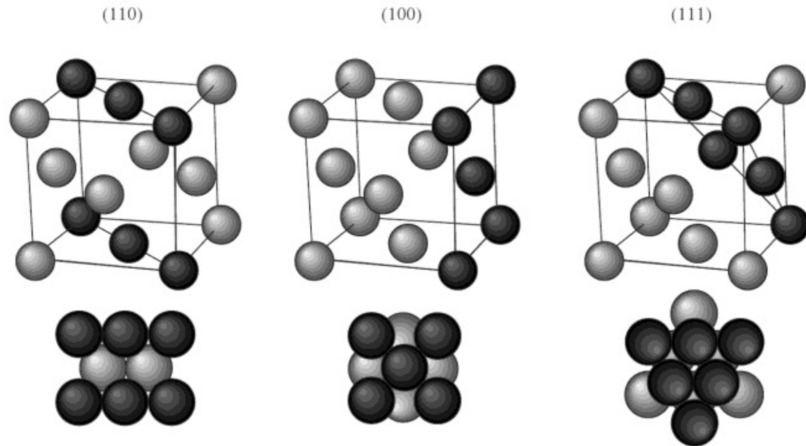


Fig. 2.4. Dense faces of a face-centered cubic crystal lattice: face (110), face (100), face (111). *Upper*: location of the face in the unit cell of the lattice. *Lower*: view from above, showing the atomic arrangement within the face itself

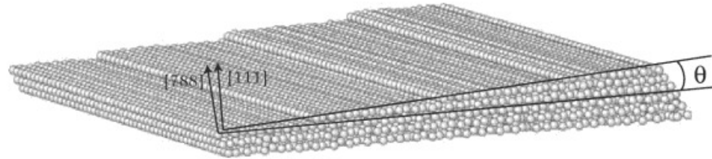


Fig. 2.5. Vicinal surface close to the (111) face, with normal (788). The angle θ is the angle by which the vicinal surface is miscut with respect to the dense face. The step density n is determined by this angle according to $n = 1/L = (\tan\theta)/h$, where h is the step height and L the width of the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

Copyrighted material

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

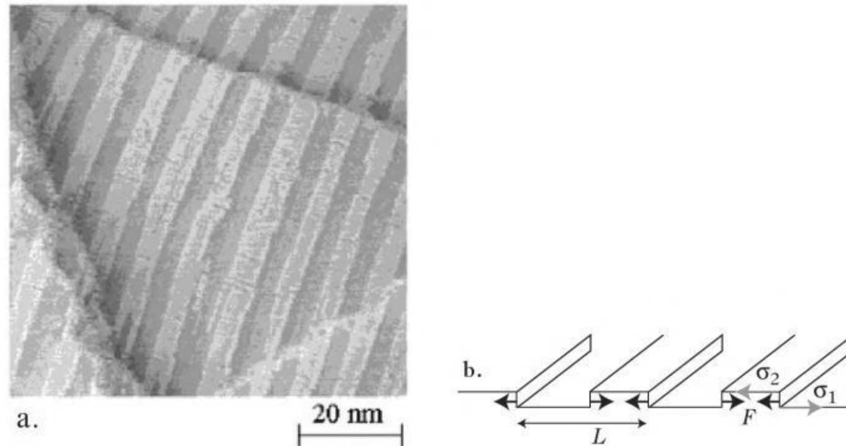


Fig. 2.6. (**a**) STM image of the copper–oxygen striped phase for oxygen coverage of 0.26 monolayers (ML) on the Cu(110) surface (1 monolayer = 1 atomic plane of the substrate, in this case copper). *Dark stripes* formed by copper–oxygen chains alternate with *brighter stripes* of bare copper. From [12] and with the kind permission of P. Zeppenfeld. (**b**) Periodic alternation (period L) of bands of oxygenated copper (stress tensor $\boldsymbol{\sigma}_1$) with bands of bare copper (stress tensor $\boldsymbol{\sigma}_2$) and the presence of forces $\boldsymbol{F} = \boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2$ along the boundary between two domains

Copyrighted material

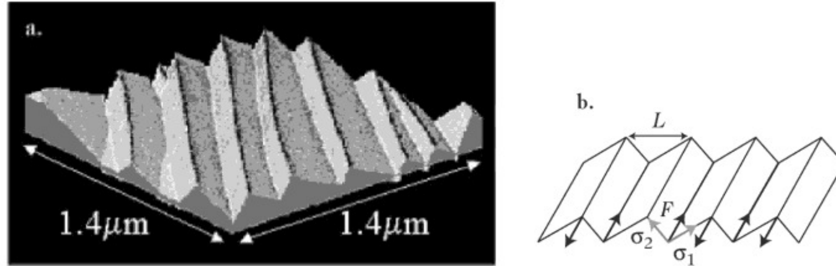


Fig. 2.7. (**a**) STM image of the faceted structure of a gold (455) vicinal surface. 3D view in which the height scale has been amplified. (The true angle between two successive facets is 174°.) The long-range order has a periodicity of 200 nm. From [16]. (**b**) Self-organisation model for a faceted surface. The period of the factory-roof morphology is L. Each facet has an intrinsic surface stress $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$. At the boundary between two facets, there are lines of force $\boldsymbol{F} = \boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2$ 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*):

Copyrighted material

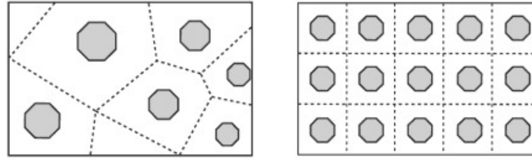


Fig. 2.10. Schematic representation of capture zones for disordered (*left*) and ordered (*right*) nucleation sites. These capture zones may result from a prestructuring of the substrate or a surface reconstruction as in Fig. 2.14

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,

Copyrighted material

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 prestuctured 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 anistropies 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

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

where F is the surface free energy and N the number of particles in the system of volume V and pressure P. Strictly speaking, μ 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 ∂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

Copyrighted material

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

where μ_0 is the chemical potential of the plane surface, Ω_0 is the atomic volume, γ 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.

Copyrighted material

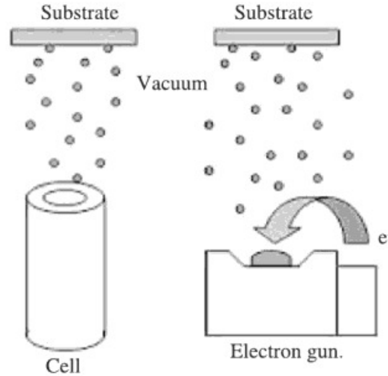


Fig. 2.11. Physical deposition methods operating in vacuum conditions. *Left*: The crucible of the effusion cell is heated by a filament to temperatures as high as 2 000°C. *Right*: The electron gun heats the material. Another method uses energy from a high-power laser beam (YAG or excimer)

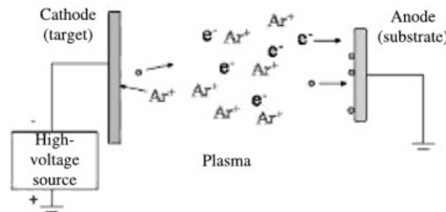


Fig. 2.12. Diode sputtering method. The surface is bombarded by a flux of high energy ions (here Ar^+) obtained from a plasma. The sputtered atoms are generally electrically neutral. Variations involve introducing a reactive element into the plasma, e.g., O_2 or N_2, or using triode setups or radiofrequencies

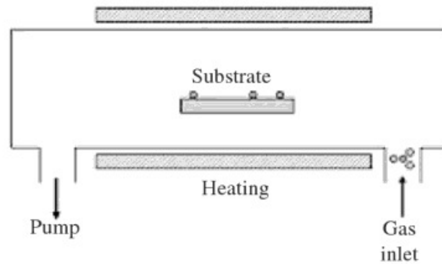


Fig. 2.13. Chemical vapour deposition (CVD), using a reactor with hot walls. Volatile compounds comprising the material to be deposited may be diluted in a carrier gas. They react on the substrate and the walls. Variations involve using cold walls and heating only the substrate, whereupon one may work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

Copyrighted material

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., SiH_4 or $AsCl_3$, 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.

Copyrighted material

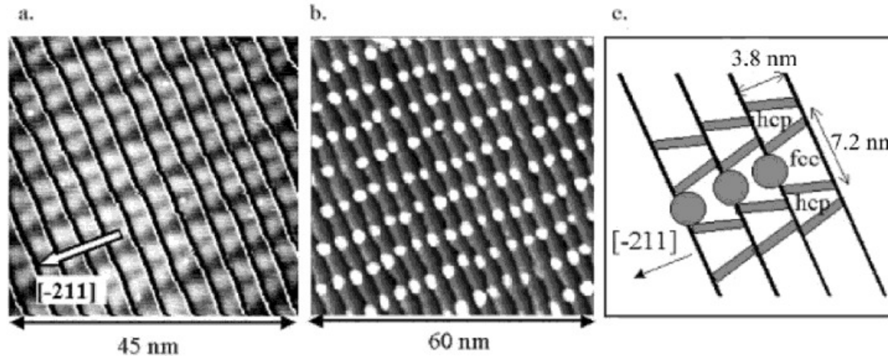


Fig. 2.15. STM images of organised growth of cobalt dots on a gold reconstructed vicinal surface. (**a**) The Au(788) substrate exhibits steps of monatomic height (0.235 nm) and brighter stacking faults perpendicular to the step edges (0.03 nm). In this image, terrace relief has been subtracted in order 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 ($a_{\mathrm{InAs}} = 6.058$ Å, $a_{\mathrm{GaAs}} = 5.653$ Å), which exhibits Stranski–Krastanov-type growth. The sawtooth 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 $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{As}$ 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

Copyrighted material

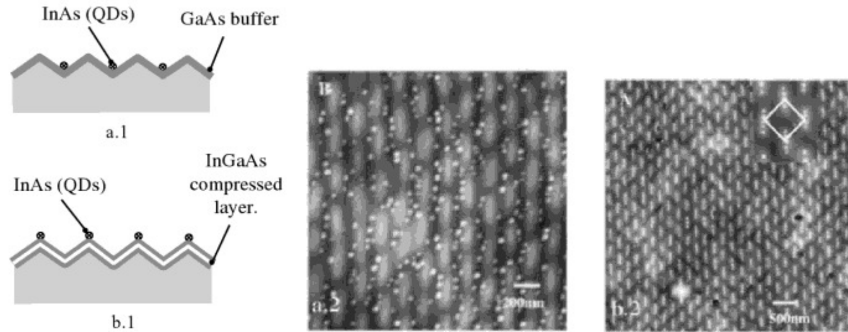


Fig. 2.16. (**a**) Quantum dots of InAs deposited on an etched GaAs(001) surface. (1) *Front view* showing the pattern. The buffer layer of GaAs is deposited on the initial etched surface to enhance the early growth stages of the InAs dots. (2) AFM image (atomic force microscopy) of quantum dots deposited 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]

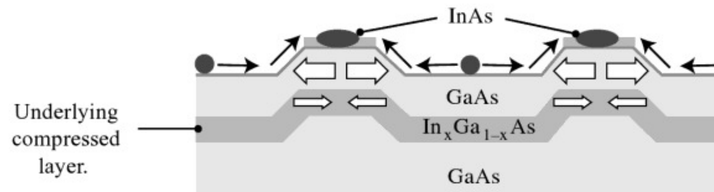


Fig. 2.17. Use of epitaxial stresses to localise quantum dots on an etched substrate [34]

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

Copyrighted material

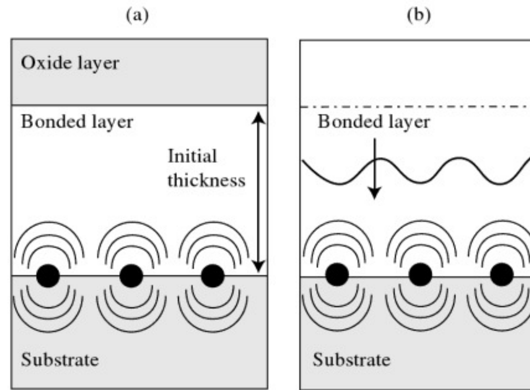


Fig. 2.18. (**a**) Molecular bonding of a thin Si film on a silicon substrate. The surface of the bonded layer is oxidised. Interface dislocations, represented by *black dots*, are the source of elastic stresses shown schematically by *curved lines*. (**b**) Chemical etching of the bonded layer of the 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 ($a_{\mathrm{Si}} = 0.5431\,\mathrm{nm}$, $a_{\mathrm{Ge}} = 0.5646\,\mathrm{nm}$). 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

Copyrighted material

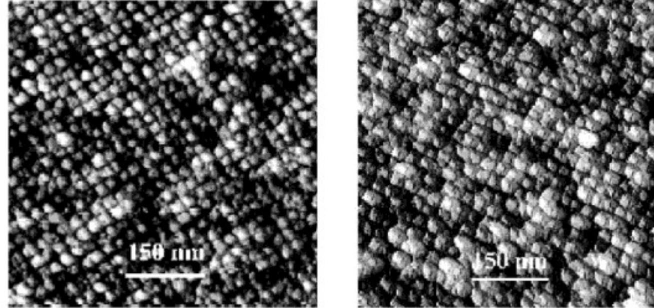


Fig. 2.19. *Left*: STM image of the surface of a bonded and etched substrate (0.88°, equivalent interval 25 nm, see Fig. 2.18). Roughness rms = 2.46 nm [38]. *Right*: Deposit of 9 Å of Ge at 450°C on a substrate obtained by chemical etching of a buried dislocation network [38]

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 prepatterning 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 prepatterning 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 Prepatterning 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 $[\overline{1}\overline{1}2]$ 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., $[11\overline{2}]$, 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

Copyrighted material

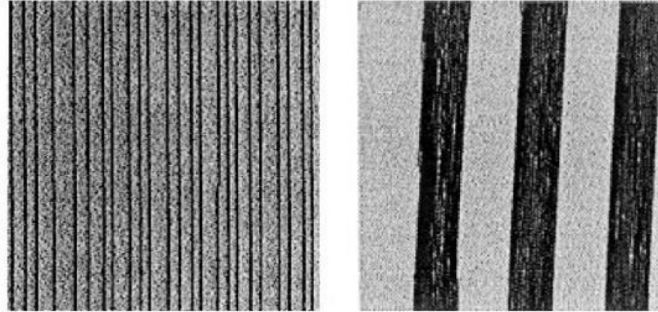


Fig. 2.20. *Left*: STM image of an Si(111) vicinal surface, miscut in the direction $[\overline{1}\overline{1}2]$. Image size $340 \times 390\,\mathrm{nm}^2$. The series of straight monatomic steps was produced by thermal treatment of the vicinal surface [40]. *Right*: STM image of an Si(111) vicinal surface, miscut in the direction $[11\overline{2}]$. Image size $240 \times 240\,\mathrm{nm}^2$. 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 $[11\overline{2}]$, which causes step bunching, an array of dimples was etched with a rotation of 30° 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

Copyrighted material

- 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., Etgens, 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)

Copyrighted material

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 Himpsel, 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 Himpsel, F.J.: Phys. Rev. Lett. **88**, 096105-1 (2002)
42. Himpsel, 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., Himpsel, F.J., and Lagally, M.G.: Phys. Rev. Lett. **85**, 5380 (2000)

Copyrighted material

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)

Copyrighted material

3

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).

Copyrighted material

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

$$I \approx V \exp(-2Kd) , \tag{3.1}$$

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.

Copyrighted material

K is of the order of $1\,\text{Å}^{-1}$. 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 μ and ν refer to the two electrodes (here, the tip and sample). Under these conditions, the tunnel current I is given by

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

where V is the applied voltage, and E_μ, E_ν are the energies associated with the wave functions ψ_μ, ψ_ν of the electrode (tip and sample) states. The main part of the calculation here is to find the matrix elements. These are given by

$$M_{\mu\nu} = \frac{\hbar}{2m} \int_S \mathrm{d}S \left(\psi_\mu^* \nabla \psi_\nu - \psi_\nu \nabla \psi_\mu^* \right) \,, \tag{3.4}$$

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

$$I \propto \frac{e^2 V}{\hbar} \rho_\mathrm{s}(r_0, E_\mathrm{F}) \rho_\mathrm{t}(E_\mathrm{F}) \,, \tag{3.5}$$

where ρ_s is the density of states of the surface measured at the tip position r_0, and ρ_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

Copyrighted material

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

$$I = \int_0^{eV} \rho_{\mathrm{s}}(r_0, E)\rho_{\mathrm{t}}(r_0, -eV + E)T(E, eV, r_0)\mathrm{d}E \,, \tag{3.6}$$

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

$$T(E, eV) = \exp\left(-\frac{2z\sqrt{m}}{\hbar}\sqrt{\frac{\Phi_{\mathrm{s}} + \Phi_{\mathrm{t}}}{2} + \frac{eV}{2} - E}\right) \,, \tag{3.7}$$

where Φ_{s} and Φ_{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 Φ.

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 Φ 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

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

Copyrighted material

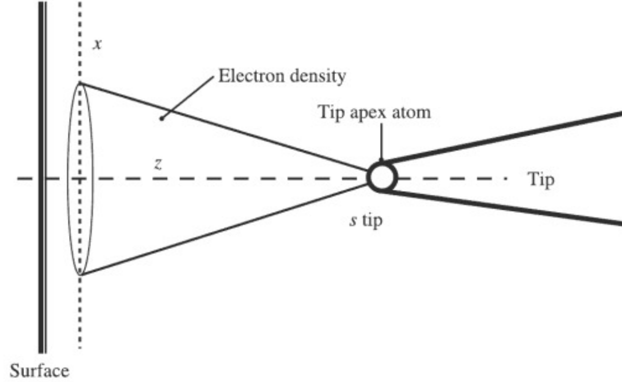


Fig. 3.2. s wave electronic density of the STM tip apex atom

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.,

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

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

$$\Delta x = \sqrt{\frac{2z}{K}} , \tag{3.10}$$

which gives the order of magnitude of the spatial resolution of the STM. With $K \approx 1\,\text{Å}^{-1}$ and z given in angstrom units, the resolution is of the order of $1.4\sqrt{z}\,\text{Å}$. 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 $1.4\sqrt{R}\,\text{Å}$. 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 Δz :

Copyrighted material

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

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

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

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 Φ of the tunnel barrier. This in turn is determined by the nature of the tip and the sample surface. Moreover, as we shall see, Φ 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 Φ of the barrier is thus defined by

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

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

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

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

Copyrighted material

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

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

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

V is the imposed bias, V_0 is a given bias [$z_0 = z(V_0)$], and $g = (\mathrm{d}I/\mathrm{d}V)/(I/V)$.

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 Φ, 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 Φ. 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

$$U(z) = \frac{1}{4\pi\varepsilon_0}\left[-\frac{e^2}{4z} - \frac{e^2}{2}\sum_{n=1}^{\infty}\left(\frac{nL}{n^2L^2 - z^2} - \frac{1}{nL}\right)\right] , \qquad (3.17)$$

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 Å 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 × 7 [9].

3.2.7 Examples

Silicon Surface

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

Copyrighted material

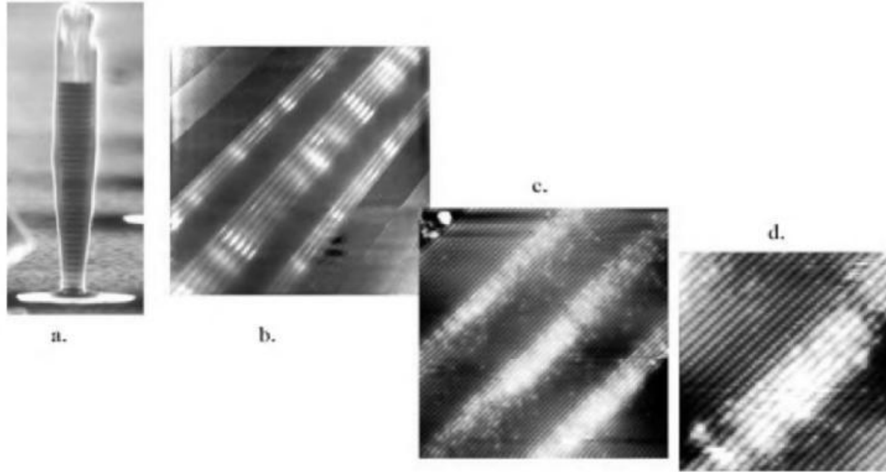


Fig. 3.5. STM images of planes of InAs quantum dots in GaAs

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 $\pm(1\text{–}3)$ 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

Copyrighted material

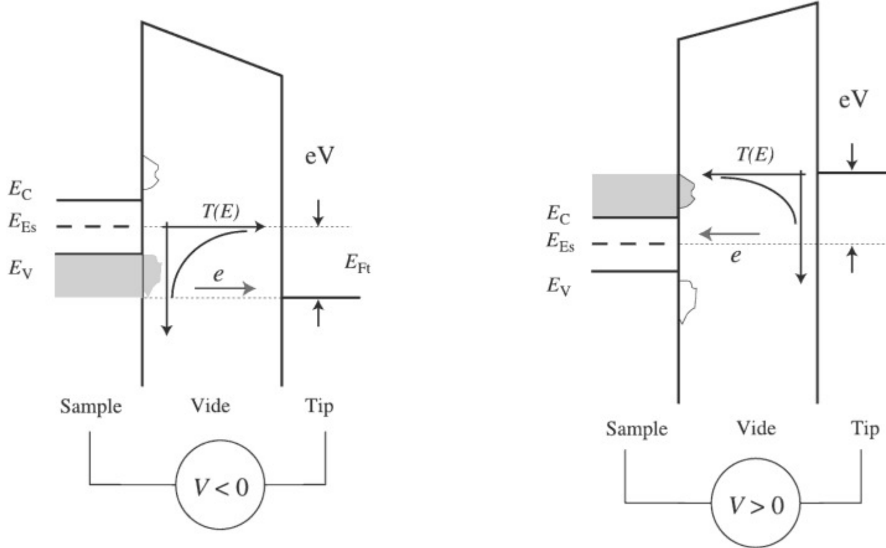


Fig. 3.6. Transmission coefficient and $I(V)$ spectroscopy

$$\frac{\mathrm{d}I/\mathrm{d}V}{I/V} \approx \frac{\rho_{\mathrm{s}}(E)}{(1/eV)\int_0^{eV}\rho_{\mathrm{s}}(E)\mathrm{d}E} . \tag{3.18}$$

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

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

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

Copyrighted material

by

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

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

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

$z(V)$ thus varies linearly with the voltage. The order of magnitude of the slope is 0.6 Å/V with $z_0 = 1\,\mathrm{nm}$ and $\Phi = 4\,\mathrm{eV}$. This slope depends on Φ 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 Γ 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 Γ 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

Copyrighted material

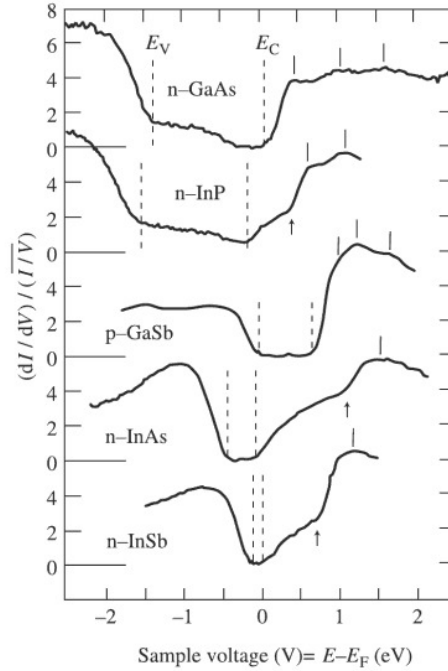


Fig. 3.7. Measuring the band gaps of III–V semiconductors. The figure shows the valence bands (negative voltage), the band gap and the conduction bands (positive voltage). For $V = 0$ V, the Fermi level of the tip is aligned with the Fermi level of the semiconductor. The maximums of the valence band and the edge Γ 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.

Copyrighted material

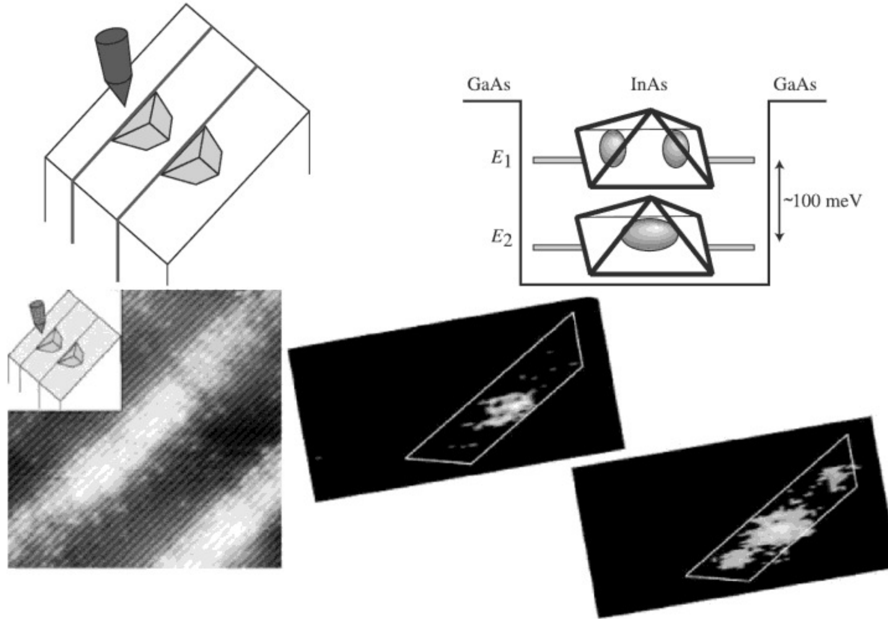


Fig. 3.8. Spectroscopy on a quantum dot

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

Copyrighted material

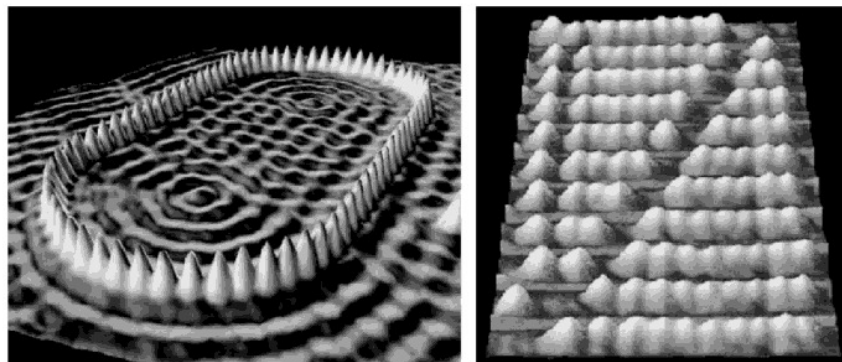


Fig. 3.10. Manipulations of atoms (*left*) and organic molecules (*right*) using an STM

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. Gimzeski 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 nm $\times$ 60 nm. Calculating the number of letters that could be written on 1 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!

Copyrighted material

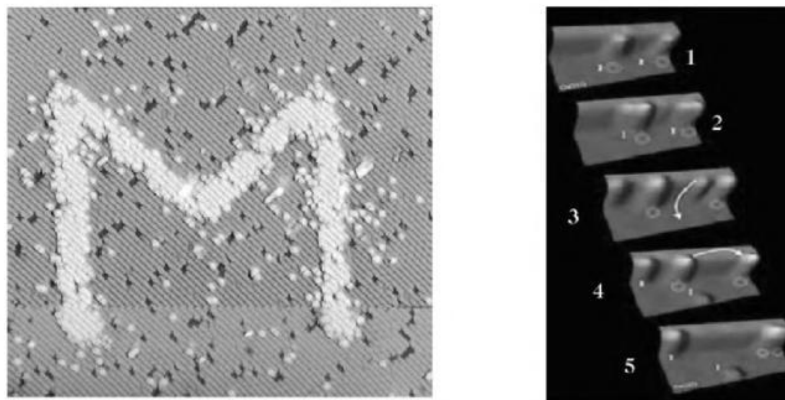


Fig. 3.11. Examples of local chemistry. *Left*: Selective dehydrogenation of a silicon surface. *Right*: Dissociation of two molecules of benzene iodide (iodobenzene) followed by formation of a diphenyl molecule

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.

Copyrighted material

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. Strocio, J.A., and Kaiser, W.J.: *Scanning Tunneling Microscopy*, Academic Press, Vol. 27 (1993)

Copyrighted material

4

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.

Copyrighted material

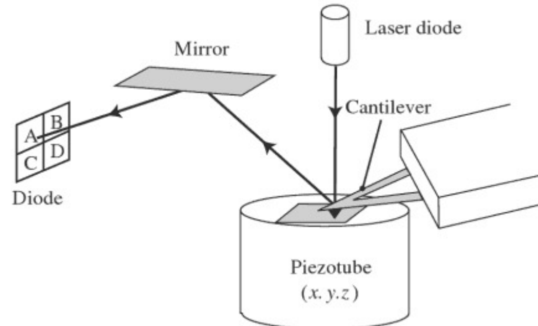


Fig. 4.1. Schematic diagram of the atomic force microscope. A piezotube displaces the sample located just below the tip carried by a cantilever. Deformations of the bolted cantilever beam are determined by measuring the displacement of the light spot from a reflected laser beam by means of a system of 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

Copyrighted material

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

Copyrighted material

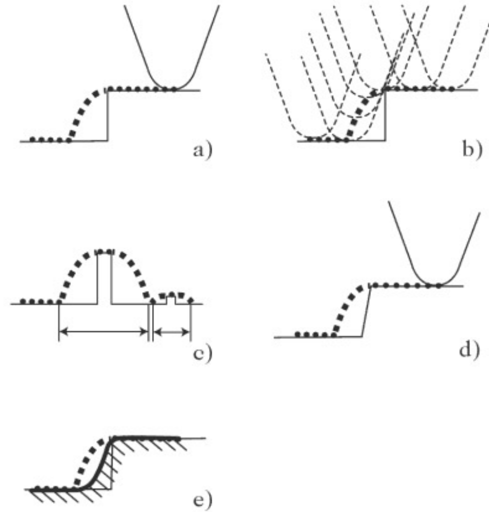


Fig. 4.5. Finite size effects due to the tip. The measured trajectory of the tip is shown by *black squares*. Successive positions of the tip as it passes over the step (**a**) are shown in (**b**). Two objects of different heights lead to different lateral extensions as shown in (**c**). A slightly 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

Copyrighted material

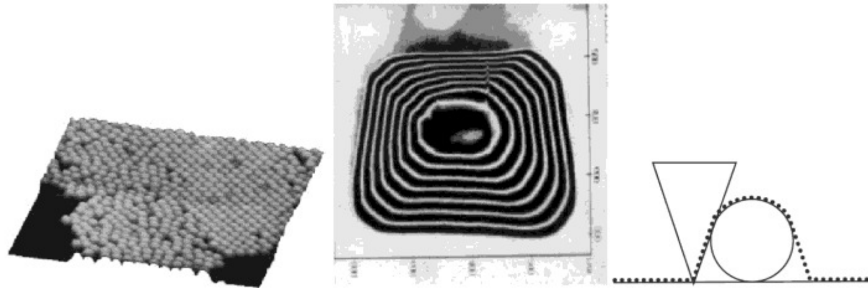


Fig. 4.6. *Left*: An island composed of a monolayer of monodispersed silicon beads. *Centre*: Single bead imaged with a pyramidal tip. The distortion caused by the tip geometry is clearly visible. *Right*: Schematic of image formation

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

Copyrighted material

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 α) 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

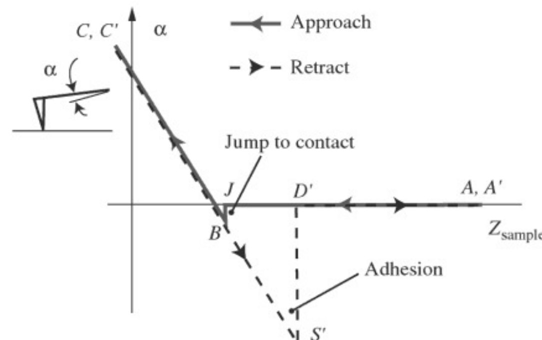


Fig. 4.7. Force curve on an ideal non-deformable material. The deflection of the cantilever beam is graphed as a function of the vertical position of the sample during an approach–retract cycle. Once contact is established at B, the deflection increases in proportion with the rise of the sample surface, 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 $S'S$)

Copyrighted material

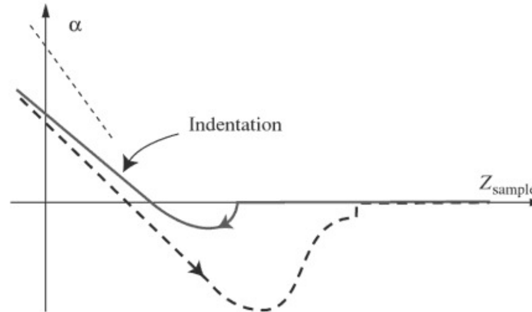


Fig. 4.8. Force curve for a deformable material. The *dashed straight line segment* shows the slope of the linear contact region which would be measured on a perfectly rigid material like the one represented in Fig. 4.7

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

Copyrighted material

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

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

where x is the position coordinate of the oscillator, which is in the present situation the displacement of the tip from its equilibrium position, ω_0 is the resonance frequency of the oscillator, γ is the amplitude of the excitation at frequency ω, β 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:

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

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 ω_0' satisfies

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

Copyrighted material

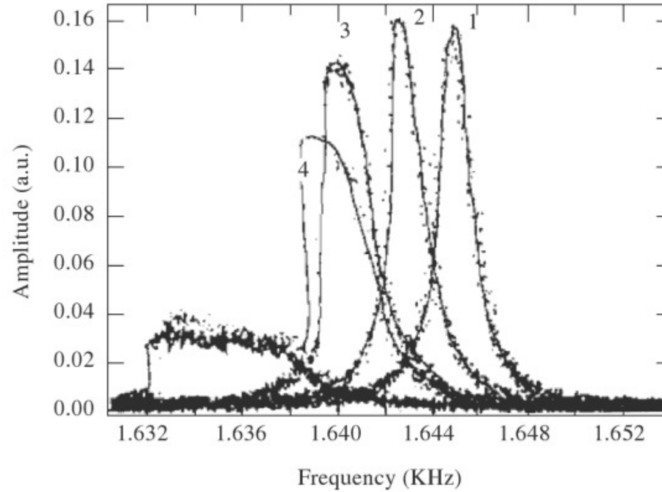


Fig. 4.10. Resonance spectra of a cantilever at various distances from an MgO surface. Curves numbered from 1 to 5 correspond to tip–sample separations of 80, 60, 50, 40 and 10 nm, respectively. Note the asymmetry of the peak at shorter distances [2]

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

Copyrighted material