Nanostructure Science and Technology

Series Editor:
David J. Lockwood, FRSC
National Research Council of Canada
Ottawa, Ontario, Canada

For further volumes:
http://www.springer.com/series/6331

Anatoli Korkin • David J. Lockwood
Editors

Nanoscale Applications for Information and Energy Systems

Springer

Editors

Anatoli Korkin
Nano and Giga Solutions, Inc.
Gilbert, Arizona, USA

David J. Lockwood
Institute for Microstructural Sciences
National Research Council of Canada
Ottawa, Ontario, Canada

ISSN 1571-5744
ISBN 978-1-4614-5015-3 ISBN 978-1-4614-5016-0 (eBook)
DOI 10.1007/978-1-4614-5016-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012949554

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Tutorial lectures given by world-renowned researchers are one of the important traditions of the *Nano and Giga Challenges* (NGC) conference series. Soon after preparations for the first forum in Moscow, Russia had begun, the organizers realized that publication of the lectures from NGC2002 would be a valuable legacy of the meeting and a significant educational resource and knowledge base for students, young researchers, and experts alike. Our first book was published by *Elsevier* and received the same title as the meeting itself—*Nano and Giga Challenges in Microelectronics* [1]. Our subsequent books based on the tutorial lectures from the NGCM2004 [2], NGC2007 [3], NGC2009 [4] conferences, and the current book from NGC2011, have been published by Springer in the *Nanostructure Science and Technology* series.

Information (electronics and photonics), renewable energy (solar systems, fuel cells, and batteries), and sensor (nano and bio) technologies have reached a new stage in their development by approaching certain engineering limits in the cost-effective improvement of current technologies. The latest miniaturization of electronic devices is approaching atomic dimensions. Interconnect bottlenecks are limiting circuit speeds, while new materials are being introduced into microelectronics manufacture at an unprecedented rate and alternative technologies to mainstream CMOS are being considered. Scaling solar energy devices based on thin films is driven by the need for cost reduction including increased efficiency of energy conversion.

Nanotechnology as the art (i.e., science and technique) of control, manipulation, and fabrication of devices with structural and functional attributes smaller than 100 nm (0.1 μm) is widely accepted as a source of potential solutions in securing future progress in information and energy technologies. It holds the capacity for massive production of high-quality nanodevices with an enormous variety of applications from computers to biosensors, from cell phone to space shuttles, and from large display screens to small electronic toys. Driven by scaling electronic devices to smaller and smaller sizes, the electronics industry has developed a set of sophisticated methods for deposition of ultrathin films with very precise composition and nanoscale lithographic patterning. Enormous investment in nanotechnology for electronics R&D from companies and governments of many developed countries now finds a "new source of revenue"—solar energy applications. Based on the same core materials—silicon and other semiconductors—photovoltaics has for a long time been a "poor relative" of micro- and optoelectronics. That is not the case any longer. The need for clean and renewable energy is becoming an imperative for the global community due to the limited resources of mineral fuel and from the growing environmental impact of our current highly wasteful use of natural resources. Moreover, the worldwide energy crisis may spark a global political crisis. Affordable access to energy and information are the linchpins on which further progress toward the solution of all global problems, and even existence of human civilization as we know it, depend.

The success of the NGC2011 conference in Russia, which resulted in the publication of this book, would have not been possible without generous support from many sponsors and research institutions. We gratefully acknowledge contributions and support from the Lomonosov Moscow State University and NT-MDT (conference hosts and co-organizers); Russian Foundation of Basic Research; Russian Nanotechnologies Corporation (RUSNANO); Springer; Intel; Sematech; our media partners STRF.RU and many other local, national, and international organizations; and individual supporters.

Gilbert, AZ, USA Anatoli Korkin
Ottawa, ON, Canada David J. Lockwood

References

1. J. Greer, A. Korkin, J. Labanowski (eds.), *Nano and Giga Challenges in Microelectronics*, (Elsevier, Amsterdam, Netherlands, 2003)
2. A. Korkin, E. Gusev, J. Labanowski, S. Luryi (eds.), *Nanotechnology for Electronic Materials and Devices* (Springer, NY, 2007)
3. A. Korkin, F. Rosei (eds.), *Nanoelectronics and Photonics: From Atoms to Materials, Devices, and Architectures* (Springer, NY, 2008)
4. A. Korkin, P. Krstic, J. Wells (eds.), *Nanotechnology for Electronics, Photonics, and Renewable Energy* (Springer, NY, 2010)

Contents

Contributors

J. Alexander NT-MDT Development Inc., Tempe, AZ, USA

S. Belikov NT-MDT Development Inc., Tempe, AZ, USA

A.M. Bratkovsky Hewlett-Packard Laboratories, Palo Alto, CA, USA

Tatyana V. Dolgova Faculty of Physics, Lomonosov Moscow State University, Moscow, Russia

Nikolai Faleev School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA

Andrey A. Fedyanin Faculty of Physics, Lomonosov Moscow State University, Moscow, Russia

Matthew P. Garrett Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Department of Physics, University of Tennessee, Knoxville, TN, USA

Rosario A. Gerhardt Georgia Institute of Technology, Atlanta, GA, USA

Dominic F. Gervasio Department of Chemical and Environmental Engineering, University of Arizona, Tucson, AZ, USA

Stephen M. Goodnick School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA

Christiana Honsberg School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA

Daniele Ielmini Dipartimento di Elettronica e Informazione and IU.NET, Politecnico di Milano, Milano (MI), Italy

Ilia N. Ivanov Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA

S. Magonov NT-MDT Development Inc., Tempe, AZ, USA

Olgierd Palusinski Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA

Maxim R. Shcherbakov Faculty of Physics, Lomonosov Moscow State University, Moscow, Russia

Chapter 1
Plasmonic Devices for Fast Optoelectronics and Enhanced Raman Sensors

A.M. Bratkovsky

Abstract Strong light confinement on the scale comparable to that of electronics components is desirable for future optical interconnects in high-performance computing systems. It would also open up new possibilities for integrated sensors with much enhanced sensitivity and selectivity. The refraction index contrast provided by group IV and III–V materials becomes insufficient for that, and one has to use metals providing much larger contrast. The dynamics of free carriers in metals is plasma like, resulting in negative dielectric constant below the plasmon frequency, $\varepsilon(\omega)<0$. This brings us into the area of plasmonics and opens up a possibility to make negative index artificial electromagnetic structures (metamaterials, NIMs) with so-called *magneto*plasmon resonances that will mimic the behavior of materials with negative permeability, $\mu(\omega)<0$, in some frequency range. We will describe various properties of NIMs and then turn over to modern Raman sensors providing single-molecule detection sensitivity. For surface-enhanced Raman probes, the achieved enhancement may reach in excess of *11 orders* of magnitude, and we describe the method of pinching gold "nanofingers" to achieve this record enhancement reproducibly.

1.1 Introduction

Silicon photonic components integrated into the state-of-the-art complementary metal–oxide–semiconductor (CMOS) technology may revolutionize modern computing platforms. CMOS-compatible silicon photonics could provide substantial size, cost, and power reduction over traditional optical communication solutions involving fiber optics and III–V semiconductors.

The proposed paradigm shift in photonics toward Si is, in general, focused on the development of four major building blocks (1) Si-based light sources,

A.M. Bratkovsky (✉)
Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA
e-mail: alex.bratkovski@gmail.com

A. Korkin and D.J. Lockwood (eds.), *Nanoscale Applications for Information and Energy Systems*, Nanostructure Science and Technology,
DOI 10.1007/978-1-4614-5016-0_1, © Springer Science+Business Media New York 2013

(2) high-index waveguides, (3) high-speed modulators, and (4) photodetectors. Metamaterials could be used in at least two of them, waveguides and modulators, offering significant miniaturization of traditional photonic components. More importantly, they can be fabricated on a large scale using traditional CMOS tools such as deposition of dielectric films, metallization, lithography, masking, and etching with the unique CMOS precision and relatively low cost. Finally, because most of these fabrication steps do not have intrinsic conflicts with Si technology, they could naturally be integrated into the CMOS environment.

Current research into artificially structured materials (also known as metamaterials) with unusual electromagnetic properties has its roots in seminal work by Lord Rayleigh in 1877 on waves in dispersive media that demonstrated the possibility of a negative group index in a medium with strong positive dispersion. In this case, the phase and group velocities have generally opposite directions, so that the medium supports *backward* propagating optical waves. Then, as was first realized by Mandelshtam [1], "negative" refraction should occur, reflecting the fact that the incident and refracted rays appear on the same side of the normal to the interface at the point of incidence.

As we will discuss below, metamaterials with simultaneous negative dielectric permittivity ε and permeability μ (negative index materials, or NIMs) have been shown theoretically to exhibit unique refractive properties [1–4] and are currently a focus of research. They enable imaging beyond a standard diffraction limit (superlensing) [5–7], transformation optics and cloaking [8–13], chiral optics [14], and demonstration of effects analogous to those in general relativity (e.g., event horizon like in "black holes") [15, 16]. The actively controllable metamaterials attract growing attention due to possible applications in computer systems, near-field devices, sensors, etc. [17, 18]. Ubiquitous materials can be used to make metamaterials, including liquid crystals [19, 20] and chalcogenides [21].

The NIM behavior at optical frequencies [22–24], which is of most interest, is essentially related to an ability of artificial periodic metal–insulator or metal–semiconductor heterostructures to support collective plasmonic modes. In particular, the so-called fishnet metal–dielectric–metal (MDM) structure supports antisymmetric plasmonic modes with a finite curl of displacement currents, curl(ωD) induced in top and bottom metallic layers ([17], cf. Fig. 1.5 in the text). Such a magnetoplasmon (antisymmetric) resonance is responsible for the effective negative $\mu < 0$ in the vicinity of the resonance. Correspondingly, one may note that the optical metamaterials belong in a wide group of plasmonic materials that are very promising for various applications requiring high field confinement due to a huge dielectric contrast between metal and dielectric [18].

1.2 Basics of Negative Refraction

The phenomenon of "negative" refraction immediately follows for a light beam entering the medium with basically oppositely directed phase ($v_{\mathrm{p}} = c/n$) and group (v_{g}) velocities, where n is the refractive index and c the velocity of light. These are related to the frequency ω of the light wave by:

$$v_{\mathrm{p}} = \omega/k, \tag{1.1}$$

$$v_{\mathrm{g}} \equiv c/n_{\mathrm{g}} = \mathrm{d}\omega/\mathrm{d}k, \tag{1.2}$$

for the phase (group) velocity $v_{\mathrm{p(g)}}$ from where we obtain the famous Rayleigh relation, recalling that $k = 2\pi n/\lambda = n\omega/c$:

$$n_{\mathrm{g}} = c(\mathrm{d}\omega/\mathrm{d}k)^{-1} = n - \lambda(\mathrm{d}n/\mathrm{d}\lambda). \tag{1.3}$$

We see immediately that it is possible to have a negative group velocity, $v_{\mathrm{g}}<0$, when $\mathrm{d}n/\mathrm{d}\lambda>n/\lambda$, i.e., the system exhibits large positive dispersion, like, for instance, at frequencies close to excitonic excitations.

The same situation may be realized in systems with strong spatial dispersion, where the phase velocity may be opposite to the group velocity in a certain frequency range. There one would observe the wave with $\vec{v}_{\mathrm{p}} \cdot \vec{v}_{\mathrm{g}}<0$ (the "backward" wave). Since $\vec{v}_{\mathrm{p}} = \hat{\vec{k}}(\omega/k)$ and $\vec{v}_{\mathrm{g}} = \partial\omega/\partial\vec{k}$, where $\hat{\vec{k}}$ is the unit vector parallel to the wave momentum $\vec{k}$, we can write down the condition for the backward wave as $\vec{k} \cdot \left(\partial\omega/\partial\vec{k}\right)<0$. The same would also be true of a hypothetical isotropic medium with both negative permittivity ε and permeability μ ("double negative" medium), analyzed by Pafomov in 1959 [25] and Veselago in 1967 [4], who established a variety of unusual properties like inverse Doppler and Cherenkov effects, in addition to negative refraction.

It is easy to show that the negative index medium supports plane waves with the opposite phase and group velocities, or $\vec{k}|| - \vec{S}$, where $\vec{S} = (c/4\pi)\left[\vec{E} \times \vec{H}\right]$ is the Poynting vector. Indeed, the Maxwell equations,

$$\begin{aligned} \frac{c}{\omega}\vec{k} \times \vec{E} &= \mu(\omega)\vec{H}, \\ \frac{c}{\omega}\vec{k} \times \vec{H} &= -\varepsilon(\omega)\vec{E}, \end{aligned} \tag{1.4}$$

give the dispersion equation and relations between the wave vector and Poynting vector:

$$\begin{aligned} k^2c^2/\omega^2 &= \varepsilon(\omega)\mu(\omega), \\ \frac{c}{4\pi}\vec{k} &= \varepsilon\frac{\vec{S}}{H^2} = \mu\frac{\vec{S}}{E^2}. \end{aligned} \tag{1.5}$$

One gets a propagating wave if and only if ε', $\mu'>0$, then $\vec{k}||\vec{S}$, which is a standard situation (forward wave, positive refraction, Fig. 1.1a), and ε', $\mu'<0$ and hence $\vec{k}|| - \vec{S}$ (backward wave, negative refraction, Fig. 1.1b).

The system is homogeneous along the interface, meaning that the momentum projection on the interface is continuous,

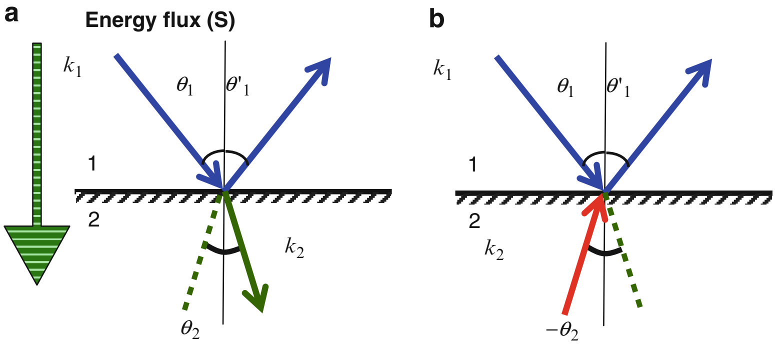


Fig. 1.1 (**a**) Normal or "positive" refraction for a second medium with positive phase velocity compared to (**b**) "negative" refraction in a system with negative phase velocity. The light source is positioned in the upper half space so that the Poynting vector S points down (*leftmost arrow*). System homogeneity along the interface dictates whether the refracted ray is "positive" (**a**) or "negative" (**b**)

$$k_1 \sin\theta_1 = k_2 \sin\theta_2, \tag{1.6}$$

which reduces to the standard Snell's law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2, \tag{1.7}$$

after we recall the linear relation between k and n [see text below Eq. (1.2)]. Usually, the correct solution corresponds to "positive" refraction, Fig. 1.1a, but one should keep in mind that there is a second solution to (1.6), $\pi - \theta_2$, or, when counted from the normal to the interface, θ_2 and $-\theta_2$. In the case of Fig. 1.1b, the phase in medium 2 advances toward the interface, as the momentum along the interface can only be conserved when one takes the "negatively" refracted ray, at the angle $-\theta_2$. From Snell's law (1.7), one would infer that this case formally corresponds to $n_2<0$, hence the notion of an NIM.

Back in 1967, Veselago [4] noticed that a parallel slab of NIM would produce an image of an object placed not farther than the thickness of the slab, l, away from it—the Veselago lens shown in Fig. 1.2.

Reflection is absent in this case, and geometric construction leads to a copy of the source formed behind the lens. Surprisingly, the copy is *exact*: "Focusing" by the NIM does not lose evanescent waves from the image, meaning that features finer than the geometric limit, $\sim\lambda$, can be resolved. It can only image objects not farther than l from the lens.

To realize a homogeneous NIM at optical frequencies, one would need to find a system with negative magnetic permeability, which is very difficult since the magnetism is a weak relativistic effect. Hence, the permeability $\mu = 1$ with very small second-order corrections in the ratio between the atomic electron velocity and the speed of light, $\propto (v/c)^2 \ll 1$ [26]. However, this argument does not apply to systems with strong spatial dispersion, as was noticed already by Mandelshtam [1],

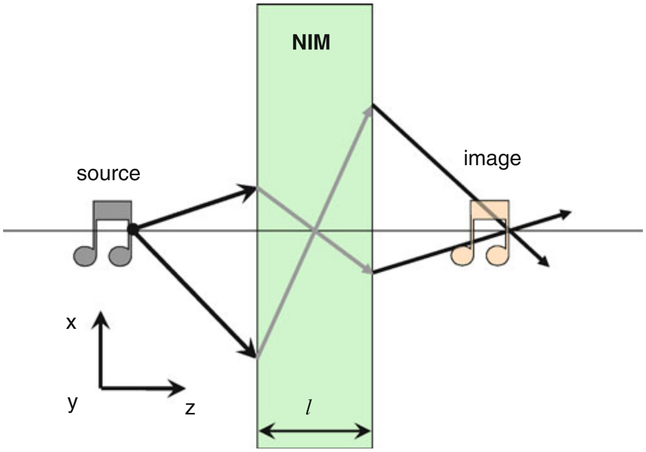


Fig. 1.2 Veselago lens made out of a parallel slab of material with $\varepsilon = \mu = -1$

where one can rather routinely get a negative group velocity, in full analogy with semiconductor crystals where one has hole bands with negative effective mass. It was realized long before photonic crystals became popular that the same situation can be arranged in two- and three-dimensional (2D and 3D) delay systems for microwave radiation, see for example Silin and Sazonov [2, 3]. They considered various 2D microwave delay lines and showed with the use of effective circuit models that is straightforward to produce various families of isofrequency curves, some of them corresponding to negative refraction in a certain frequency range. The same case was discussed decades later by Notomi for the case of photonic crystals in [27]. Indeed, depending on symmetry and position in the Brillouin zone, some photonic bands correspond to backward waves $\omega = \omega_{\mathrm{r}}(\vec{k})$, where $\vec{k}$ is the wave momentum defined in the first Brillouin zone of the system and r is the number of the band. Some of the isofrequency bands have negative curvature, $\vec{k} \cdot \vec{v}_{\mathrm{g}} < 0$, where $\vec{v}_{\mathrm{g}} = \vec{\nabla}_{\vec{k}}\omega$ is the group velocity. The case of photonic bands in a crystal with negative curvature ("holes") corresponds to negative refraction [27]. One may also easily identify the possible flat regions of photonic bands (extended van Hove singularities) with "slow light" behavior, similar to heavy carrier bands in semiconductors, see Fig. 1.3.

Typical photonic crystal dispersion is shown in Fig. 1.3b [27], showing pockets of "hole" like behavior in the areas marked 1, 1′. The correspondence with the semiconductor band dispersion is pretty direct (Fig.1.3b): "hole" like pockets correspond to negative group velocity and, consequently, to negative refraction.

1.3 Superlens Effect

Remarkably, the NIM lens makes it possible to exceed the geometric limit for imaging with the Veselago lens. The usual resolution of spatial features of the imaged object is limited by the wavelength λ used for imaging. The classical

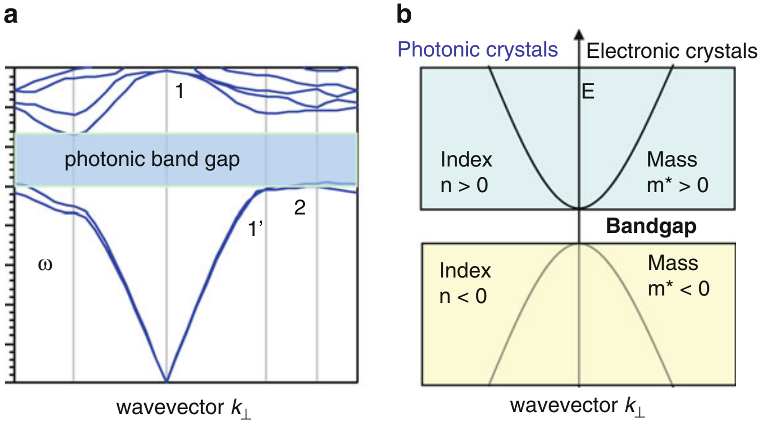


Fig. 1.3 (**a**) Typical band structure of a photonic crystal with photon dispersion and (**b**) its schematic compared with the electron dispersion in a typical semiconductor crystal. The photon dispersion in areas 1,1′ is analogous to hole dispersion (**b**), while flat band region 2 corresponds to a small group velocity, i.e., "slow light." "Conserved" $k_\perp$ indicates projection of the momentum of the photon on the interfacial plane (cf. Fig. 1.1)

explanation consists in the observation that the rays incident at grazing angles at the lens (cf. Fig. 1.2) cannot propagate in the lens but form evanescent waves exponentially decaying into the lens material and producing an exponentially small contribution to the image. This leads to information being lost from the source, thus limiting the resolution. Paradoxically, at first sight, this does not happen in the Veselago lens, since the evanescent waves are enhanced inside the NIM, as was discovered by Pendry [5]. The physical reason for this unusual behavior [5, 28] lies in a well-known fact of NIM supporting surface plasmon polaritons (SPPs) [26]. In fact, the NIM slab works as an electromagnetic resonator, with SPPs being pumped by the radiation coming from the source/object itself [29], while even slight deviations from ideal conditions severely limits the resolution [29] as do the losses [5, 30], so the "perfect" superlens is not really possible. A surface plasmon corresponds to surface currents (electron density oscillations, mainly out of the plane, in plane ones are overdamped) that reradiate power from the back surface and restore the exact field intensity corresponding to the evanescent waves that would have been lost otherwise. The effect was indeed observed with the simplest "superlens," a silver thin film [6, 7]. Silver is not an NIM per se; it has negative $\varepsilon < 0$, but its permeability $\mu = 1$, as in any metal *below* the plasmon frequency that lies in the optical range [see Eq. (1.12)]. However, this does not matter, since for p-polarization (*H*-field in plane of the slab) the magnetic component is conserved automatically and μ drops out. Fang et al. [6] showed that the resolution in this case can reach about $\lambda/6$, i.e., well into sub-λ regime. Since then, even finer relative resolution as a fraction of the wavelength λ has been achieved with various metallic and SiC superlenses for far-infrared and microwave regions. Overall, all things related to metamaterials become much easier when one moves to longer wavelengths, while anything in optical domain faces stiff challenges.

“Superlensing” is generally a weak effect limited by the losses and surface imperfections of the lens [29]. To see the effect of an NIM slab enhancing the evanescent waves, the core aspect of superlensing, one *must* solve the full Maxwell equations, since the effect is absent in the geometric optics approximation by definition. Consider, for simplicity, the (only nonzero) H_y component of the p-polarized incident field (Fig. 1.2). Then:

$$\begin{gathered} H_y^{(\mathrm{evan})}(z,x) = \sum_{k_x} H_{k_x} \mathrm{e}^{\mathrm{i}k_x x - \kappa z}, \\ \kappa = \sqrt{k_x^2 - \varepsilon\mu \frac{\omega^2}{c^2}}, k_x > \varepsilon\mu \frac{\omega}{c}, \end{gathered} \tag{1.8}$$

where k_x is the wavevector for the ray incident at the slab at a grazing angle and, hence, decaying exponentially along the z-axis. The fields in the source and the image are related as:

$$H_y^{(\mathrm{image})} = t\mathrm{e}^{-\kappa l} H_y^{(\mathrm{source})}, \tag{1.9}$$

where t is the transmission coefficient, which is exponentially large, totally compensating the exponential decay of the field in vacuum in the ideal case $\varepsilon = -1$, where $t = \mathrm{e}^{\kappa l} > 1(!)$. This is the superlens effect: the evanescent part of field in the image appears to be exactly the same as in the source.

The superlens effect in a more general situation deviating from the ideal one corresponds to $\varepsilon = -1 + \delta,\ \delta \ll 1$, where the transmission deviates from the pure exponential law [29]:

$$t = \mathrm{e}^{\kappa l} \left(1 - \frac{\delta^2}{2} \mathrm{e}^{\kappa l} \sinh \kappa l \right)^{-1}. \tag{1.10}$$

The spatial size of features that can be resolved can be related to the maximal wavevector k_{c}, where the denominator in (1.10) remains close to unity (so that near perfect restoration of the evanescent field takes place). With the use of (1.8):

$$\Delta x \sim 2\pi/k_{\mathrm{c}} \approx \frac{\lambda}{\sqrt{1 + \frac{\lambda^2}{4\pi^2 l^2} \ln^2 \frac{2}{|\delta|}}} < \lambda. \tag{1.11}$$

We see that in order to have an appreciably better resolution than λ, one should use slabs much thinner than the wavelength, $l \ll \lambda/2\pi$ (near field), and a material close to an ideal one. In the experiment [6], the operating wavelength was 365 nm and the Ag layer thickness was very small, $l = 35$ nm $\ll \lambda$, so that $\lambda/2\pi l = 1.7$, which would suggest $\delta \leq 0.1$.

1.4 Fishnet NIM

Besides subwavelength resolution, a compact fast optical modulator with picosecond switching time that could be used, for example, as a subwavelength coupler switch between two optical ports is very interesting if it can operate at the communication wavelength of $\lambda = 1{,}550$ nm. Since the unit cell of the corresponding metamaterial should be much smaller than λ, one is looking at structures with a period of ~100–300 nm, i.e., the optical applications of NIM necessarily require the use of nanostructures. Split-ring resonators that could produce the required negative permittivity and permeability do not work at the desired optical frequencies since the response of metals at these frequencies is quite different from that in the microwave regime. Namely, frequency dispersion becomes very strong, or in other words the response becomes *plasmonic*. In noble metals of most interest for NIM applications [26],

$$\varepsilon(\omega) = \varepsilon_{\mathrm{d}} - \omega_{\mathrm{p}}^2/[\omega(\omega + \mathrm{i}\gamma)], \tag{1.12}$$

where ω_{p} (γ) is the plasmonic frequency (decay rate) tabulated in, e.g., [31, 32], and ε_{d} the background dielectric constant related to interband transitions.

The split-ring resonators (SRR) geometry that works at microwave frequencies does not apply in the optical regime and is difficult to fabricate, but one can use a metal–dielectric hole array structure that we call the fishnet [22]. In short, the fishnet structure (hole array in a metal film stack) can be viewed as a composite material with the linear arrays of quantum metallic wires providing an $\varepsilon < 0$, and the perpendicular wider metallic nanowires that support an antisymmetric plasmon with a nonvanishing curl of the displacement current in bottom and top wires producing an effective $\mu < 0$ [17]. Below, we describe a design of a "fishnet" NIM operating at 1,550 nm and analyze possible ways of countering the losses by adding gain medium. We finish with a brief discussion of the results of optical modulation and its physical origins.

We have designed fishnet NIMs like Ag/SiO$_2$/Ag and Ag/Si/Ag [17, 33], the latter with a semiconductor spacer layer specifically for the purpose of optical modulation by full-scale Finite Difference Time Domain (FDTD) simulations [34, 35] (some details discussed below). We have found that losses in our structures are rather hefty and they should be countered by an added gain medium, as we discuss below.

The fabrication combined electron beam lithography (EBL), nanoimprint lithography (NIL), and a lift-off process [33]. First, a Si NIL mold with "fishnet" patterns was made by EBL and reactive ion etching. Second, a transfer layer and a liquid UV-curable NIL resist (imagining) layer were coated onto a glass substrate by spin coating. Then, the fishnet patterns were imprinted into the imaging layer using a UV-curable NIL process with the Si mold. After the patterns were transferred into the transfer layer by residue-layer and transfer-layer RIE etchings, the Ag/SiO$_2$/Ag stack was deposited using electron beam (E-beam) evaporation. Finally, a lift-off process was used to remove the transfer layer and imaging layer to leave the stack fishnet structure directly on the glass substrate (more details are given in [33], see Fig. 1.4 illustrating the fabrication steps).

In our design, we have used two 25-nm thick Ag films with a 35-nm thick SiO_2 spacer layer in between. The widths of the metallic "wires" that composed the fishnet have been 100 nm × 300 nm. While the smallest feature in our structure was about 100 nm, which sounds large compared to our prior applications, our studies showed that NIM fabrication required very high precision for both lateral and vertical dimensions. The calculated position of the magnetoplasmon (antisymmetric plasmon between top and bottom metallic layers, see above) that corresponds to a dip in the real part of the refractive index n', λ_r, shifts by about 10 nm with either a 1.3-nm Ag or a 0.7-nm SiO_2 layer thickness variation, or a 1.6-nm line width change.

The sensitivity of the lateral dimension is especially challenging for the state-of-the-art fabrication. Normally, EBL is used to pattern NIMs, and the feature size variation of 10 nm from run to run is rather common; therefore, it is hard to fabricate NIMs with better than 60 nm accuracy in position of the resonance λ. Note that the modeling could not predict the performance of the fishnet exactly, because all of the material parameters we used were taken from the corresponding bulk materials, while the important characteristics like the relaxation time in nanostructures might deviate strongly from their bulk values due to surface scattering, grains, etc. Therefore, revision of the design based on feedback from experimental data is necessary. The nonrepeatability associated with EBL, even as small as 10 nm, makes this trial-and-error process very difficult and laborious. We have solved those problems by using NIL technology. Even though the NIL molds were fabricated by EBL, which has about 10 nm feature size uncertainty, subsequent patterning by NIL has excellent repeatability. Hence, we have developed the tuning-in strategy of fabricating NIMs for an accurate working wavelength [33] (1) we fabricate the first-generation NIMs using NIL and the lift-off processes and measure the offset from the desired wavelength (i.e., 1.55 μm); then (2) we fabricate the second-generation NIMs using the same processes with the same NIL mold but tune the thickness of stack to compensate the offset, as shown in Fig. 1.5.

1.5 Fishnet NIM at 1,550 nm

Initially, we have performed an extensive FDTD modeling of the resonant fishnet structures in order to design the system that will have a magnetoplasmon resonance and effective negative index at 1.55 μm [35]. Our design led to a Ag 33 nm/Si 80 nm/Ag 33 nm structure [17] with the scattering characteristics displayed in Figs. 1.4 and 1.5. The unit cell size and the lateral sizes of intersecting nets of metallic (nano)wires are shown in Fig. 1.5. The calculated transmission,

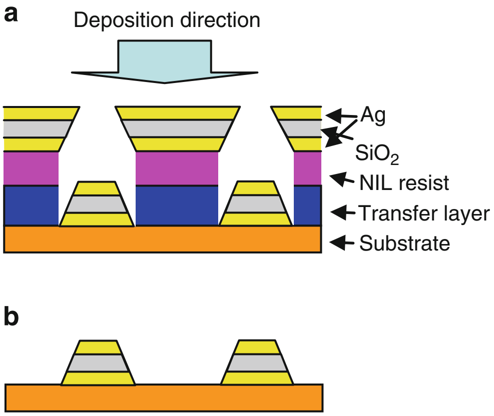


Fig. 1.4 Schematic of deposition and lift-off steps after nanoimprinting the structure into NIL resist during fishnet fabrication. (**a**) The opening in the resist structure is getting smaller during Ag and SiO$_2$ deposition by E-beam evaporation. (**b**) Fishnet with nonvertical sidewalls left on the substrate after the lift-off process

reflectance, and absorption characteristics exhibit a very well-defined resonance at 1.55 μm (Fig. 1.6).

Linear transmission and reflection spectra have been measured in order to characterize the sample. The measurements have been carried out with a Nd: YAG laser/optical parametric system generating 20 ps pulses tunable in the entire near-IR range [17]. For transmission measurement, the input beam was normal to the surface with polarization parallel to the thin Ag wires (cf. Fig. 1.5). For reflection measurement, the same polarized beam was tilted by 10° to the normal to the surface. The phase difference was also measured in transmission and reflection of the two beams polarized orthogonally to each other, parallel to the thin and thick Ag wires, respectively. To study the pump-induced change of transmission and reflection, we used the pump/probe method using Q-switched YAG:Nd^{3+} doubled output at wavelength of 532 nm as the pump pulses and the time-delayed tunable IR pulses from Optical Parametric Oscillator (OPO) system as the probe. The pump is expected to excite carriers Si and Ag, and we wanted to find out which ones would mainly change the optical responses of the fishnet.

To prove that the designed metamaterial indeed supports backward waves, which is a necessary condition for negative refraction, we have also estimated the transmission phase that appears to be *negative* in a wide range of frequencies around the resonance, see Fig. 1.6c. This confirms that the fishnet indeed supports the backwards waves.

The real part of the effective index, n_{eff}, appears to be negative at the 1.55 μm resonance in a wide range that spans about 100 nm, see Fig. 1.7. This result is in accordance with the estimates showing that both magnetic permeability and dielectric permittivity are negative in a wide region near the resonance, see Fig. 1.8.

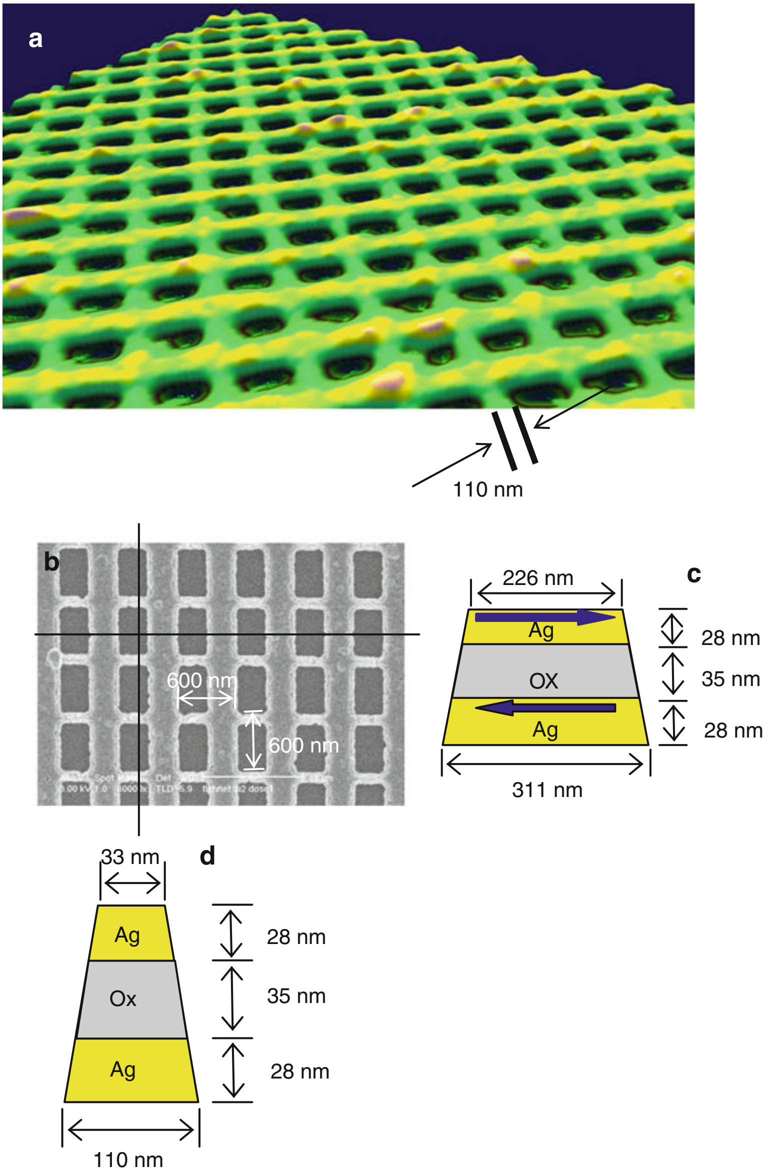


Fig. 1.5 (**a**) AFM image of the "passive" fishnet Ag/SiO_2/Ag with the dip of refractive index $n' = -1.7$ at $\lambda = 1{,}560$ nm. (**b**) SEM image of the fishnet. The dimensions are marked on the SEM image and the schematics of the cross section. (**c**) and (**d**) The cross sections of the fishnet across the wide and narrow set of wires, respectively. The *arrows* in panel (**c**) show the directions of flow of the displacement current J in wide wires when the magnetoplasmon is excited, so that curl$\vec{J} \neq 0$, and system exhibits an effective $\mu < 0$

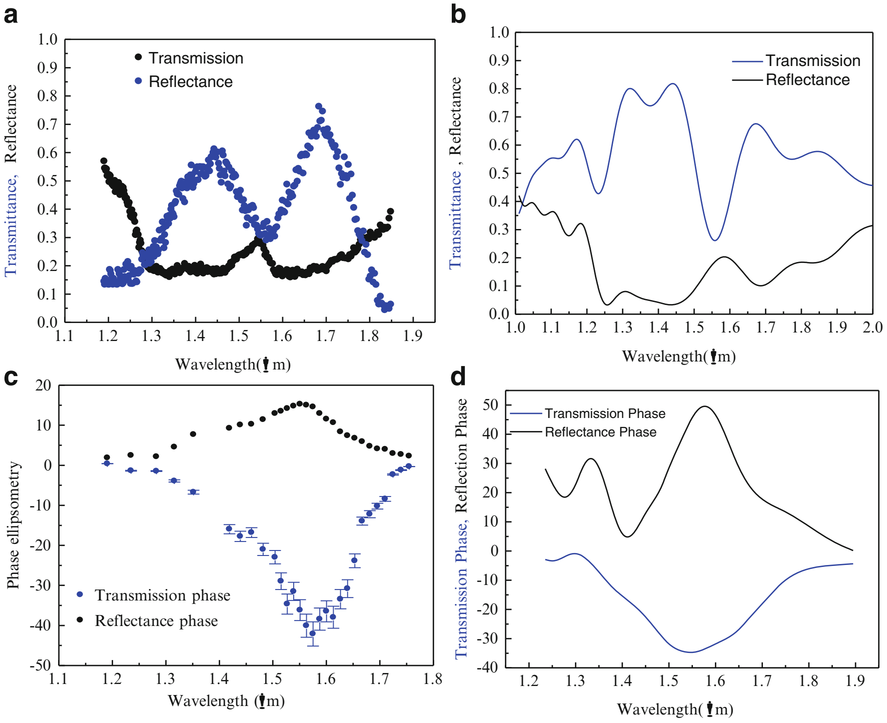


Fig. 1.6 Transmittance, reflectance, and absorption of the fishnet structure Ag 33 nm/Si 80 nm/Ag 33 nm. The data for transmission and reflectance (**a**) compared with FDTD simulations (**b**). Experimental data (**c**) and theoretical prediction (**d**) of the phase anisotropy

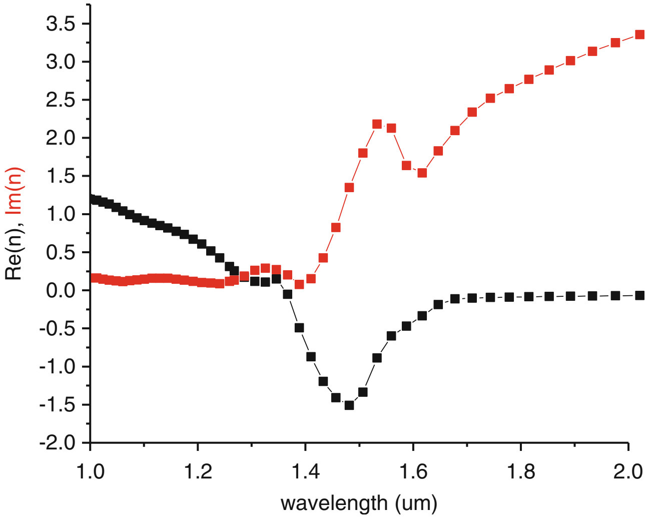


Fig. 1.7 Effective refractive index of the resonant fishnet structure (real and imaginary parts)

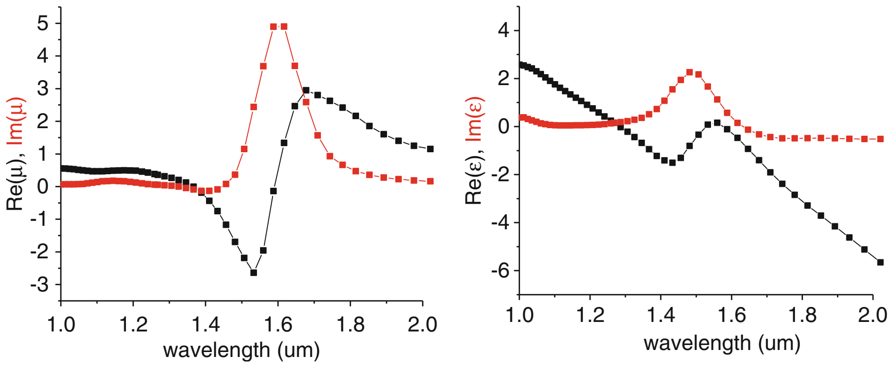


Fig. 1.8 Magnetic permeability and dielectric permittivity of the fishnet structure; their real (*black*) and imaginary (*red*) parts

1.6 Index Modulation in Fishnet NIM

A similar fishnet but with a Si spacer layer instead of SiO_2 spacer has been used by us in optical modulation experiments [36]. The results have indeed shown that it is possible to strongly change the effective index of refraction by pumping the Si spacer layer, creating a large number of photocarriers there (density of optical carriers $\sim 10^{18}\mathrm{cm}^{-3}$), thus increasing the conductivity of the spacer layer substantially and changing the strength of the magnetoplasmon resonance that corresponds to the coupling between the top and bottom layers, as explained above. Obviously, by making the spacer layer much more conducting, one tends to weaken the magnetoplasmon resonance, Fig. 1.5c, substantially or wipe it out totally.

We show in Fig. 1.9a, b the phase anisotropy and the transmittance of the fishnet sample with and without the pump. The unoptimized fishnet had the resonance at ~1.7 μm in agreement with the theoretical prediction. The pump fluence was 320 μJ/cm^2. It is seen that the pump induces a decrease of ~50% in the peak magnitude of the magnetic resonance in the transmittance spectrum and a red shift of 15 ± 2 nm. The observed phase difference, Fig. 1.9a, is mainly due to the effect of the magnetic resonance seen only by the waves with the polarization of the magnetic field vector along the thick Ag wires, see Fig. 1.5c. Without the pump, it reaches $-38°$ in transmission and 60° in reflection at the resonance. With the pump, the values change to $-25°$ and 42°, respectively, following the changes in the resonance.

The real and imaginary parts of the effective refractive index, n, can then be deduced from the experimental data in Fig. 1.9 using the method of [37], with the results shown in Fig. 1.9c. At the magnetic resonance, $\mathrm{Re}(n_{\mathrm{eff}})$ for the fishnet exhibits a dip with the value at the minimum $\mathrm{Re}(n_{\mathrm{eff}}) = -2.4$. With the pump

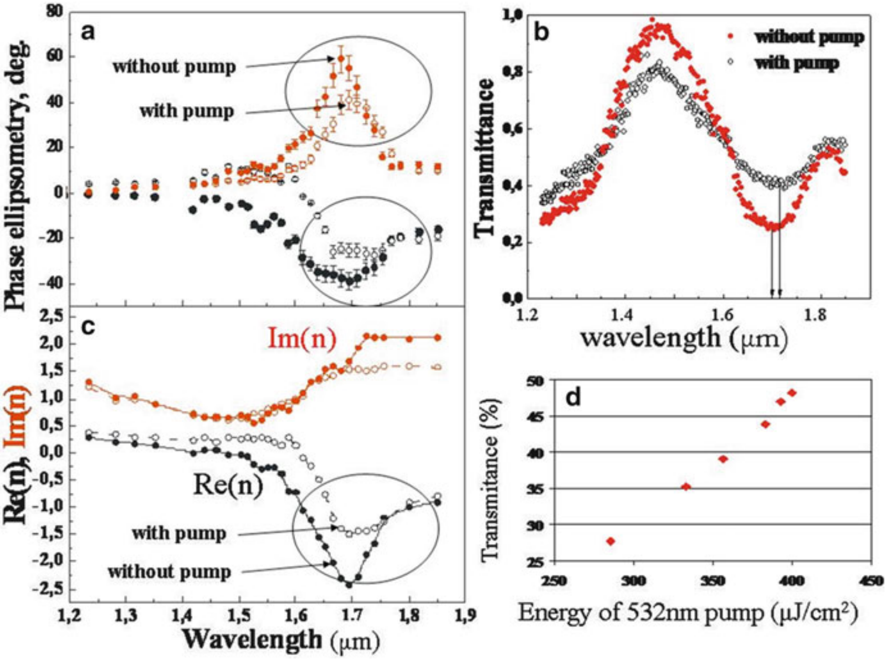


Fig. 1.9 Phase anisotropy (**a**), transmittance (**b**), real Re(n), and imaginary Im(n) parts of the effective index (**c**), and the increase of transmittance with 532 nm pump fluence (**d**). In (**a**), phase difference spectra for transmitted and reflected light without the pump (*solid red* and *black dots*, respectively) and with the pump (*open red* and *black dots*, respectively) are shown. In panel (**b**), the transmittance without and with pump is shown by *solid* and *open circles*, respectively. In (**c**), real (*black*) and imaginary (*red*) parts of the effective index are shown. Data without pump are shown by *solid symbols*, with pump by *open symbols*

reducing the resonance strength, it changes to $\mathrm{Re}(n_{\mathrm{eff}}) = -1.5$, which is a huge change. We also show the measured transmittance at the magnetic resonance as a function of the pump fluence in Fig. 1.9d. The linear relation indicates that the effect is due to the proportional increase of pump absorption in the structure.

Pumping the sample (1) produces free carriers in Si and Ag and their relaxation also leads to (2) heating of the sample. Both processes could modify the optical constants of the materials and hence the optical response of the metamaterial, but carrier relaxation is expected to be much faster than heating. In our pump/probe measurement, we measured a set of transmission spectra at various delay times between pump and probe and observed the relaxation of the induced changes on the resonance structure. Pump-induced change of transmission at the resonance peak is shown in Fig. 1.10a as a function of the probe time delay for a pump fluence of 320 μJ/cm^2. For comparison, we also display in Fig. 1.10b the cross-correlation trace of our pump and probe pulses obtained from sum-frequency generation in a barium borate crystal.

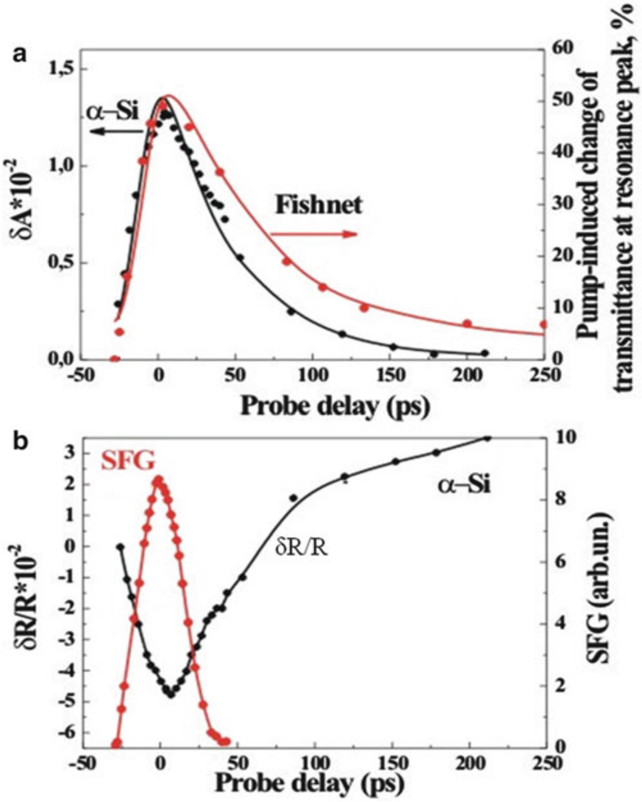


Fig. 1.10 (**a**) Pump-induced transmission change at the magnetoplasmon resonance of the fishnet (*red dots*) and pump-induced absorption variation from an amorphous Si film (*black dots*) as a function of the probe time delay for a pump fluence of 320 μJ/cm^2. (**b**) Cross-correlation trace of pump and probe pulses obtained from sum-frequency generation in a barium borate crystal (*red dots*) and the time-resolved pump-induced reflectance change from an amorphous Si film (*black dots*)

One can fit the experimental data for the fishnet in Fig. 1.10a with the following expression:

$$S \propto \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \mathrm{d}t_1 \mathrm{d}t_2 \theta(t_2 - t_1) \cdot \left(\mathrm{e}^{-a(t_2-t_1)} + A\right) \cdot \mathrm{e}^{-\frac{2(t_2-\tau)^2}{w^2}} \mathrm{e}^{-\frac{2t_1{}^2}{w^2}}. \tag{1.13}$$

Here, $\theta(t_2 - t_1)$ is the step function, the exponential decay term $\exp[-\alpha(t_2 - t_1)]$ and the constant A describe, respectively, the effects of carrier relaxation and heating on modulation; the Gaussian functions represent the pump and probe pulse profiles that reproduce the cross-correlation trace in Fig. 1.10b with pulse width $w = 19$ ps and $\tau = 58$ ps the time delay between pump and probe pulses, see Fig.1.10a.

We assume that the pump-induced modulation comes mainly from the photo-excitation of carriers in the α-Si layer. In that case, the observed fast relaxation would reflect the carrier relaxation in the Si layer. A similar pump/probe measurement has been carried out on an 80-nm thick α-Si film alone, without top and bottom Ag layers. The pump-induced absorption and reflectance changes versus time for the α-Si film are shown in Fig. 1.10a and 1.10b, respectively. The initial dip in reflectance results from a pump-induced reduction of the refractive index and can be associated with pump-induced free carriers in α-Si. The change in the absorption (Fig. 1.10a) is determined from simultaneous measurements of reflectance and transmittance. The decay of absorption changes can be fit by a single exponential with a relaxation time of 50 ps (Fig. 1.10a), characteristic of carrier relaxation in α-Si [38, 39]. The fact that the pump-induced modulation of our fishnet sample has a decay close to that of a free α-Si film indicates that free carrier excitation and relaxation is indeed the dominant mechanism responsible for the modulation, while the excitations in the Ag wires appear not to be important. The tail observed at a long probe delay time, on the other hand, must have resulted from a thermal modulation as the excited carriers relaxed and released their energy to heat up the sample.

To further confirm that carrier excitation in α-Si is the dominant mechanism underlying the observed pump-induced modulation of the fishnet structure, we deduced a relatively small maximum pump-induced refractive index change of $\Delta n_{\mathrm{Si}} = \Delta n_{\mathrm{Si}}' + \Delta n_{\mathrm{Si}}'' = -0.055 \pm 0.01 + i(0.02 \pm 0.005)$ from the pump/probe measurement of the α-Si film (without silver layers) for a pump fluence of 320 μJ/cm^2. The imaginary part of the index is due to finite conductivity of photo-induced carriers that we estimated to be about $\sigma_{\mathrm{ph}} = 1.3 \times 10^{13}\ \mathrm{s}^{-1}$. The FDTD calculation of the effective refractive index of the fishnet structure with the changes in α-Si refractive index $\Delta n_{\mathrm{Si}} = -0.055 + i0.008$ and $\Delta n_{\mathrm{Si}} = -0.055 + i0.048$ shows a red shift of the magnetoplasmon resonance [the dip in $\mathrm{Re}(n_{\mathrm{eff}})$] of about 5 and 30 nm and a reduction of the resonant amplitude by 30% and 70%, respectively. The results are in fair agreement with the experimental observation shown in Fig. 1.9. The control pump/probe measurement on a fishnet structure with the SiO_2 spacer layer replacing Si has been made too, and we found that the maximum pump-induced change of transmittance with 320 μJ/cm^2 is less than 5%. Because silica does not absorb at the pump frequency, the effect, if any, would have come from the pump effect on Ag wires. This again indicates that modulation of the fishnet structure by excitation of the Ag wires is not effective.

Summarizing this part, our data and FDTD simulations have both demonstrated photo-induced modulation of the effective negative refractive index of a Ag/Si/Ag fishnet structure. A pump with fluence of 320 μJ/cm^2 at a visible wavelength has changed the effective refractive index of a Ag/Si/Ag fishnet negative index structure at the resonance from $n_{\mathrm{eff}} = -2.4 + i1.7$ to $n_{\mathrm{eff}} = -1.5 + i1.5$. Photo-induced carriers in the α-Si spacer are responsible for the modulation. Generally, it is characterized by dynamic response time of 58 ps governed by the carrier relaxation time in α-Si.

1.7 Fast Modulation in a Picosecond Range

As described above, we have observed a modulation with a relaxation time of the order of 50 ps in the setup that allowed a resolution of the dynamics down to about 10 ps [17]. It may be related to either fast recombination at the surface states in the present thin film (80 nm of Si spacer) or bulk trapping of charges. Since the spacer material is likely amorphous, the mobility and diffusion coefficients for photo-induced carriers $D = 0.1–1\ \mathrm{cm}^2/\mathrm{s}$. Then, the characteristic diffusion time to cover the distance $l \leq 40$ nm and hit the surface would be $\tau_\mathrm{s} \sim l^2/D$~20–150 ps. Since the spacer material may be partially crystalline with a somewhat higher mobility, the relaxation time may be even faster ~1 ps. In addition, l is comparable to the mean intertrap distance in our Si spacer. Therefore, the trapping of the carriers may also lead to a ~1 ps fast relaxation time. The slow tail that we observed earlier, ~50 ps, is related to interband transitions and other possible relaxation mechanisms involving deep traps [38, 39].

More recently, we have used femtosecond pulses to drive the system and observed much faster initial relaxation on the order of just 1 ps. This would open up completely new possibilities for extremely fast optical switching. We would need more data that will allow us to gain more insight into the fast relaxation processes in our samples and to better control them.

1.8 Fishnet with Gain Medium

It is clear from the results [41] that the losses in the metal–dielectric hole array stack (fishnet) have been pretty large. This is a fundamental problem with plasmonic devices that necessarily use metals supporting SPPs. The field of SPPs penetrates into the metal leading to significant heating losses.

One way of mitigating the losses would be to add a gain medium that should help to recover the signal. In the present work, we have mounted the fishnet (Fig. 1.11a, b) on top of stack of five InGaAsP quantum wells (Fig. 1.11c) to check the effect of a gain medium substrate. The InGaAsP multiple quantum wells is the gain medium of choice that can be pumped optically at 1,550 nm. To this end, we have designed with the use of the FDTD method [34] an optical metamaterial that comprised a fishnet on top of five quantum InGaAsP wells with a buffer layer of 3–20 nm InP between QWs and the fishnet, see Fig. 1.11. The change of the substrate to InGaAsP with a refractive index of $n_\mathrm{S} \approx 3.2$ red shifts the resonance, this is why the geometry of the structure should be readjusted in comparison with a fishnet on a standard glass substrate. We have calculated the transmittance and reflectance spectra of the fishnet structure on two different substrates: one just InP and the other a stack of five InGaAsP/InGaAs quantum wells.

The samples were designed to have the magnetic resonance at 1.55 μm. The fishnet stacks were 0.6 nm Ge/20 nm Ag/25 nm $\mathrm{SiO_2}$/0.6 nm Ge/20 nm Ag. The Ge

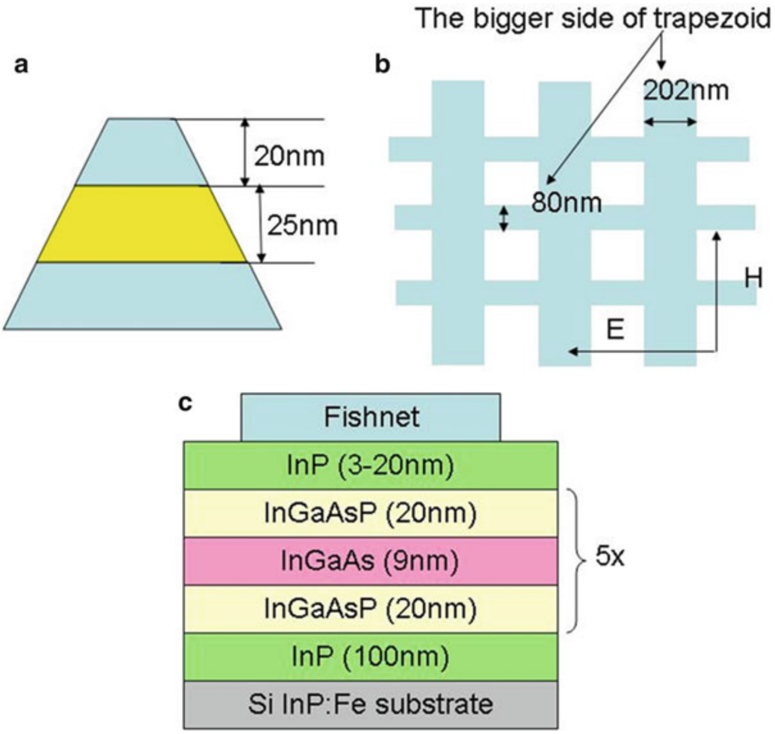


Fig. 1.11 Schematic of designed fishnet structure Ag 20 nm/SiO_2 25 nm/Ag 20 nm: side view (**a**) and top view with base dimensions shown (**b**). InP gain substrate with the Ag/SiO_2/Ag fishnet on *top* (**c**)

layers were used to create smooth Ag films [33]. Smooth Ag surfaces are essential for reducing loss in “fishnet” metal–dielectric structures. However, it is more difficult to achieve a smooth Ag film on the quantum well substrate directly, because Ag does not wet the InP surface. By adding a 0.6 nm Ge film as the wetting layer, both surface roughness and line edge roughness of the fabricated “fishnet” structures improved dramatically [33]. The structures and dimensions of the samples are sketched in Fig. 1.11.

The pumped MQW substrate has been simulated with the gain parameter $\sigma = 3{,}000\mathrm{cm}^{-1}$ [40]. As one can see from Fig. 1.12, adding this considerable gain into the substrate produced a modest effect on both transmission and reflection. The position of the resonance did not change, but the change in the figure of merit, Re(n)/Im(n), has been predicted to be small. Preliminary measurements performed at UC Berkeley by David Cho and Ron Shen seem to conform to these expectations, see discussion in [41].

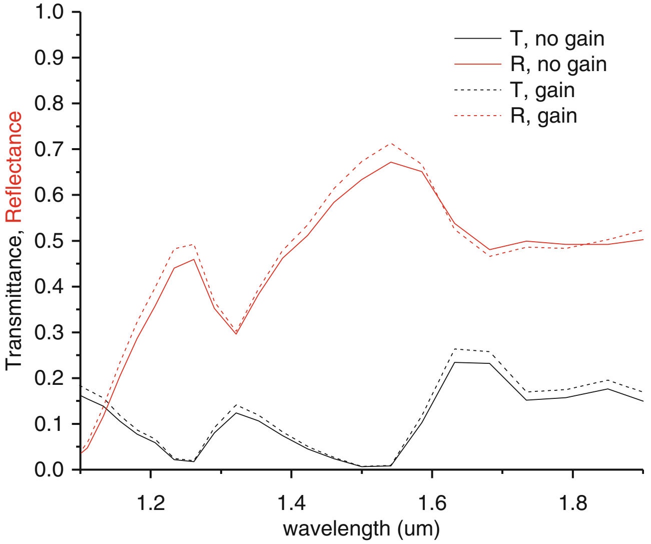


Fig. 1.12 Transmission and reflection: FDTD results for the fishnet without and with 3,000 cm^{-1} gain (MQW stack with 20 nm InP buffer layer)

1.8.1 Germanium NIM Fishnet: Effect of Narrow Band Spacer

As discussed above, the Ag–Si–Ag fishnet with thin Si spacer layer (about 80 nm) exhibits a large modulation depth and a fast fall-off time, which is below 1 ps [41]. Indeed, in spite of a tiny thickness of the whole fishnet, about 100 nm in total, the depth of the modulation is approaching 50% due to resonant character of response. At the same time, small thickness may be a decisive factor in fast response related to very fast relaxation of photocarriers at traps and interfaces between Si spacer and metal Ag fishnet.

It would be interesting to tailor properties of the spacer layers, especially try those with narrower band compared to Si, since one may be able to inject more (photo)carriers and have shorter relaxation times to produce faster switching with better modulation ratio. To this end, we have studied theoretically and experimentally the possibility to use Germanium as the spacer material. Ge has considerably lower bandgap (0.7 eV) compared to Silicon (1.1 eV). Germanium is highly technological CMOS-compatible material and would be preferable if it would be possible to make an NIM with it. There is, of course, a large family of narrow band materials that would be interesting to explore for the novel NIM applications. Since we are interested in communication wavelengths around $\lambda_c = 1{,}550$ nm $= 0.8$ eV, the main question that we address here is whether the direct interband absorption with the threshold sitting at exactly the same energy (direct band gap is $E_0 = 0.8$ eV, while the indirect band gap is $E_g = 0.67$ eV) would not destroy negative index behavior.

Fortunately, the NIM behavior is possible in germanium fishnet around the communication wavelength λ_c, as we show below. For a proper account for Ge absorption, we used very accurate Adachi model for the dielectric response, incorporating it into the FDTD method [34]. Then, we have optimized a geometry of the fishnet structure starting from rescaling the Si fishnet to a higher refractive index of Ge to have magnetoplasmonic resonance at around λ_c.

The optical absorption in Ge as a function of an angular frequency ω is described by $\varepsilon_2(\omega)$:

$$\varepsilon_2(\omega) = \left[\left(4e^2\hbar^2\right)/\left(\pi m^2\omega^2\right)\right]P^2 J_{cv}(\omega), \tag{1.14}$$

where $P = <c|p|v>$ is the momentum matrix element between conduction (c) and valence (v) bands, $J_{cv}(\omega)$ the joint density of states. The direct gap transitions contribute the most to $\varepsilon_2(\omega)$ and since we are interested in the behavior at the wavelengths $\lambda > 1{,}300$ nm ($\hbar\omega < 1$ eV), in practice, we need to basically account for transitions in the center of the Brillouin zone with transition energies $E_0 = 0.8$ eV, $E_0 + \Delta_0 = 1.09$ eV, where $\Delta_0 = 0.29$ eV is the spin–orbit split-off of the valence band at the Γ-point [42]. Assuming that the corresponding bands are parabolic, one obtains a simple golden rule expression as:

$$\varepsilon_2(\omega) = A(\hbar\omega)^{-2} \times \left[(\hbar\omega - E_0)^{1/2}\theta_{E_0} + \frac{1}{2}(\hbar\omega - E_0 - \Delta_0)^{1/2}\theta_{E_0+\Delta_0}\right], \tag{1.15}$$

where $A = 2.70$ eV$^{3/2}$ is the constant expressed through the combined density of state mass m^* and P^2, $\theta_\mathrm{E} = 1$, when $\hbar\omega > E$, and zero otherwise. The real part $\varepsilon_1(\omega)$ then readily follows from the Kramers–Kronig relation. Below the threshold for direct transition in the region $1.55 < \lambda < 1.85$ μm, Fig. 1.13, the absorption is dominated by indirect phonon-assisted processes. This is a higher order process compared to the direct transitions:

$$\varepsilon_2^{\mathrm{ind}}(\omega) \propto (\hbar\omega)^{-2}\left(\hbar\omega - E_\mathrm{g} \pm \hbar\omega_q\right)^2, \tag{1.16}$$

where $\hbar\omega_q < 40$ meV is the energy of the phonon taking part in the transition, and it is much smaller than the contribution of the direct transitions, see data in [43]. Since the plasmon resonance and NIM region is rather broad because of metal losses, Fig. 1.14, it is not sensitive to the small contribution from indirect transitions and broadening of the absorption edge in Ge fishnet. Those effects have been neglected, therefore.

In the finite difference time domain that we used, one needs to propagate the field in time, and for the displacement field D one has:

$$\begin{aligned}\vec{D}(\vec{r},t) &= \int_0^t \mathrm{d}t'\varepsilon(t-t')\vec{E}(\vec{r},t') \\ &= \varepsilon_0\varepsilon_\infty\vec{E}(\vec{r},t) + \int_0^t \mathrm{d}t'\chi(t-t')\vec{E}(\vec{r},t'),\end{aligned} \tag{1.17}$$

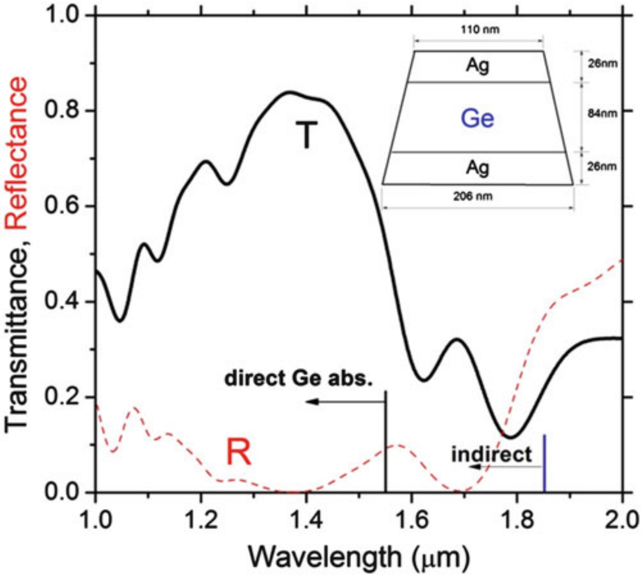


Fig. 1.13 The transmission and reflection coefficients for Ge fishnet. The region of direct and indirect (phonon assisted) interband absorption in Ge are marked by *arrows*. The *inset* shows one of the vertical cross sections of the fishnet structure

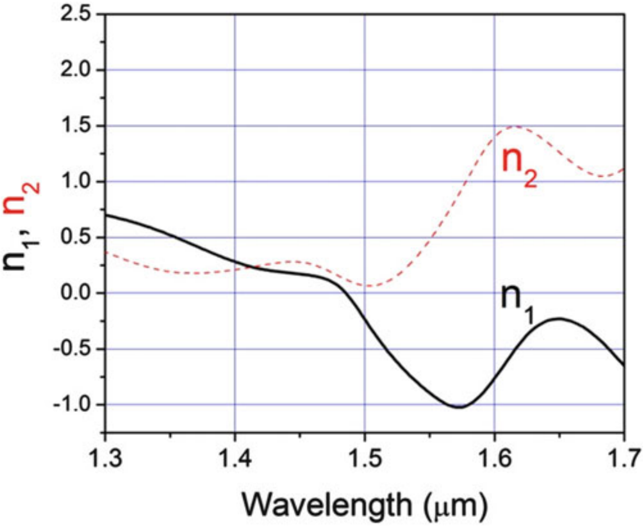


Fig. 1.14 Real (n_1) and imaginary (n_2) parts of the effective index of the fishnet structure estimated from the transmission data in Fig. 1.13. Negative index behavior is observed in a wide range around 1,550 nm wavelength

where ε_∞ is the high frequency dielectric constant of Ge, ε_0 the dielectric constant of vacuum, and one needs to know the response function $\chi(t)$ in all preceding moments in time (we assume that the field is zero at the time $t = 0$). One can construct $\chi(t)$ for the Adachi's model (1.15) by using the Kramers–Kronig relations to find $\varepsilon_1(\omega)$ [42], then $\chi(\omega)$ and, finally, $\chi(t)$ by numerical Fourier transform on a time grid needed for the FDTD simulations. Then, one can use it in the Maxwell's equation for curl$\boldsymbol{H}$ to find the update for the $E(r,t)$ through its values at all preceding moments in time. Such an algorithm together with the Adachi's model produces results that cannot be reproduced with oversimplified Lorentzian approximation for $\varepsilon(\omega)$, as the test study of the laser pulse propagation in GaAs in [44] has shown.

For silver layers, we have used the standard Drude model [34]:

$$\varepsilon_{\mathrm{Ag}}(\omega) = \varepsilon_\infty^{\mathrm{Ag}} - \frac{\omega_{\mathrm{p}}^2}{\omega(\omega + i\Gamma)}, \tag{1.18}$$

where parameters $\varepsilon_\infty^{\mathrm{Ag}}$, $\omega_{\mathrm{p}} = 1.182 \times 10^4$ THz, and $\Gamma = 1.213 \times 10^2$ THz are the same as in prior study [17], chosen to fit the tabulated dielectric permittivity for silver [32].

The designed Ge fishnet on a glass substrate consisted of two 26-nm thick Ag metallic layers separated by a 84-nm thick Ge layer and perforated by a periodic array of holes (Fig. 1.13, inset). The period of the resulting network of metallic wires is 310 nm along the wires in both directions. The width of the Ag wires along the two perpendicular directions is 206 nm and 75 nm for the bottom layer, respectively and is approximately 40% smaller for the top layer. The latter (trapezoidal) vertical cross section takes into account our fabrication procedure involving nanoimprint and liftoff that produces slanted sidewalls instead of the vertical ones, as discussed above.

The transmittance (Fig. 1.13) has a dip (peak in reflectance) near $\lambda = 1.55$ μm and it corresponds to the negative index behavior in the same region, see Fig. 1.14. One would notice that the fishnet is pretty transparent at shorter wavelengths $\lambda < 1.55$ μm with median reflection coefficient at around 10% while transmittance reaches 85%. The refractive index $n = n_1 + in_2$ estimated from the transmission coefficients is shown in Fig. 1.14. The real part $n_1 \approx -1$ in the vicinity of $\lambda = 1.55$ μm as designed. This corresponds to a broad magnetoplasmon resonance at the communication wavelength. The resonance is accompanied by the peak in absorption at longer wavelengths so that the figure of merit is $n_1/n_2 = 1.4$. This is not due to Ge interband absorption since it kicks in at shorter wavelengths but apparently due to metal losses. Since the metal losses are generic, one has to bear pretty large losses while working with metal–semiconductor negative index metamaterials, including those with Ge or Si spacer layers.

Summarizing, we have shown on the basis of accurate Adachi model for Ge absorption used in our FDTD code that the Ge absorption edge does not destroy the negative index behavior at important 1,550 nm communication wavelength. Since the band gap (0.67 eV) is substantially smaller than the one in Si (1.12 eV), the hope

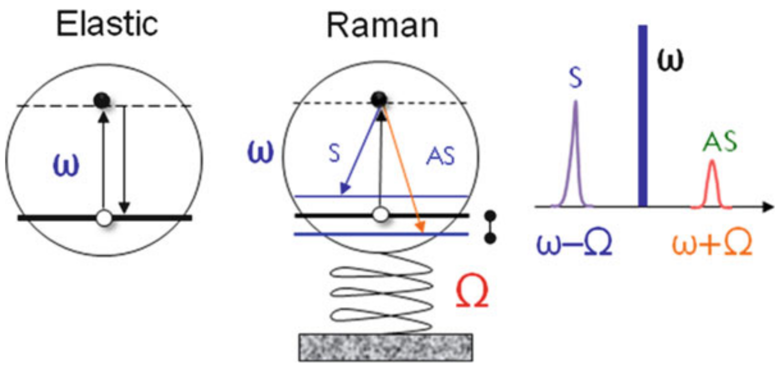


Fig. 1.15 Schematic of the elastic (*left*) and inelastic (Raman, *center*) processes of light scattering on a molecule. The *thick solid lines* indicate the ground state, *broken line* the virtual or real excited electronic state. Raman processes generate: Stocks (S) and anti-Stocks (AS) side lines shifted from the central line ω to $\omega \mp \Omega$, where Ω is the frequency of a particular excited molecular vibrational mode

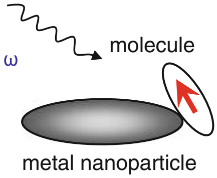


Fig. 1.16 Local field enhancement near small metallic object, e.g., a noble metal nanoparticle, leads to enhanced Raman scattering. Since this is an inelastic (higher order) process, Raman cross section may be enhanced by many orders of magnitude

is that would lead to a deeper depth and even faster (sub-ps) modulation compared to Si fishnet [41]. The preliminary data for optically pumped Ge fishnet show that this may indeed be the case. In the future, one would like to use the naturally supplied top and bottom electrodes of the fishnet (silver layers) for electrical modulation of the device.

1.8.2 Surface-Enhanced Raman Spectroscopy

Inelastic (Raman) scattering of molecules provides a unique fingerprint of the species and thus provides a uniquely specific probe in comparison with elastic scattering, Fig. 1.15. Surface-enhanced Raman scattering (SERS), Fig. 1.16, was discovered in 1974 [45] and given an adequate interpretation in 1977 [46, 47], see an extensive discussion in e.g., [48]. The SERS effect is apparently due to very strong local field enhancement near metal nanoparticles or rough metallic surface [48–50]. Below, we shall describe the SERS produced by proximity of the molecules to individual nanoparticles, like shown in Fig. 1.16, and squeezed between ensembles of nanoparticles in a controlled way. In the latter case, the enhancement may exceed eleven orders of magnitude (!).

Raman scattering is due to the dipole electronic transition in a molecule being modulated by much slower (compared to the frequency ω of the electronic transition) atomic vibrations with a typical frequency $\Omega \ll \omega$ (usually, for molecular vibrations have the energy in the range $\hbar\Omega = 1$–100 meV, while for an incident near-infrared or visible laser light $\hbar\omega = 1$–2 eV). This is a classical explanation of Raman inelastic scattering that obviously remains valid in the quantum mechanical picture as well [51].

The induced dipole moment d on the molecule is proportional to the local field E,

$$d_i = \alpha_{ij} E_j, \tag{1.19}$$

where α is generally the second rank tensor, $i,j = x,y,z$ the Cartesian components (will e omitted below for simplicity). Since $E = E_0 \cos(\omega t)$, then $d = \alpha\, E_0 \cos(\omega t)$, and the polarizability is modulated by the molecule vibrations, $\alpha = \alpha + \alpha_1 \cos \Omega t$, substituting these expressions into Eq. (1.13) yields:

$$d = \alpha_0 E_0 \cos \omega t + \tfrac{1}{2}\alpha_1 E_0 [\cos(\omega - \Omega)t + \cos(\omega - \Omega)t]. \tag{1.20}$$

This result suggests that the scattered wave will contain unshifted (central) Rayleigh line, the Stokes line at ω–Ω and the anti-Stokes line at $\omega + \Omega$.

Specific Raman process should be allowed symmetry selection rules, i.e., the corresponding matrix element:

$$M_{fi} = \left\langle \psi_f \middle| \alpha_{1,ij} \middle| \psi_i \right\rangle, \tag{1.21}$$

should be nonzero, and this takes place only when the product of irreducible representations of ψ_{i}, the induced dipole ($\alpha_{1,ij} E_0$), and ψ_{f} comprises a totally symmetric representation. Note that the Cartesian components of the polarizability tensor α_{1xx}, α_{1xy} transfer as the corresponding vectors products xx, xy, ... One should check the representations of the vibrational initial and final states of the molecule to see if the corresponding product in (1.21) contains fully symmetric representation to allow for Raman excitation [51].

1.9 Silver Octopods Nanoparticles

Metallic nanoparticles are the focus of research due to multitude of possible applications, especially due to possibilities of their surface functionalization, uses as plasmonic “nanorulers” [52], local plasmon related field amplification [53], etc. The latter is particularly interesting for employing them in the SERS studies [45].

Since the surface plasmonic resonances (SPRs) are extremely size- and shape sensitive, the particles with the shapes other than solid sphere are very interesting for SERS applications. Due to their high local curvature, presence of sharp points, protrusions, and large aspect ratio, the particles like cubes, disks, nanorods, nanoshells, prisms, cuboctahedra, bipyramids, and multipods can effectively amplify the electric field in their vicinity [54–60], and those shapes have been demonstrated experimentally, see e.g., review [61]. Special attention has been paid to metallic nanostars [36, 61–64], and very recently a highly symmetrical type of nanostars, the silver octopods have been synthesized [65]. By exposing octahedron-shaped Ag nanoparticles to an etchant with the preferential etching along the [100] direction, the authors obtained *isolated* multiarmed octopod structures maintaining the initial cubic O_h symmetry. It can be viewed as a solid sphere with eight cylinders protruding along the [111]-like crystallographic directions, Fig. 1.1 (insets a,b). As will be clear below, this star shape is of particular interest for Raman scattering, this is why we will discuss it in detail.

To understand the plasmonic features of the nanostars, we investigated their scattering properties using the discrete dipole approximation (DDA) [66, 67]. In this approach, the object of interest is represented as an array of polarizable mini-spheres on a cubic grid with a lattice period a. The period a is supposed to be taken much smaller than the wavelength of the incident light ($a \ll \lambda_0$), so that the polarizable mini-spheres could be treated in the quasistatic approximation. Each sphere feels the field of the incident beam $\mathbf{E}_{\text{inc}}$ and the fields generated by all other spheres, as described by a system of linear equations for polarization of all mini-spheres: for i-th mini-sphere $\mathbf{p}_i = \alpha_i \mathbf{E}(\{\mathbf{p}_i\})$, where $\mathbf{E}(\{\mathbf{p}_i\})$ is the standard retarded dipole field produced by all other mini-spheres with index $j \neq i$. The polarizability of the i-th mini-sphere is given by the standard Lorenz–Lorentz expression $\alpha_i^{-1} = r_i^{-3}(\varepsilon_i + 2)(\varepsilon_i - 1)^{-1} - 2ik^3/3$ for the sphere with radius r_i, with the last complex term giving the radiative correction [68].

The octopods/nanostars have been specified by three geometrical parameters: the core radius R, the cylinder radius r, and by $L = R + h$, where h is the length of the cylinders, Fig. 1.17 (inset b). The top area of the cylinders was chosen to be not flat but rounded as a segment of the sphere with the radius L. One can also describe the geometry of the octopods by two dimensionless parameters L/R, r/R and an effective radius $a_{\text{eff}} = (3V/4\pi)^{1/3}$ expressed via the total volume of the particle V. The grid spacing a was taken as 2 nm to facilitate convergent results.

We begin by demonstrating that the present geometrical model of the nanostars indeed works and is consistent with the data. Shown in red in Fig. 1.17 is the experimental scattering spectra obtained for a single particle encased in a SiO_2 bead [65]. The optimal fit was obtained with the following set of geometrical parameters: $a_{\text{eff}} = 70.5$ nm, $L/R = 1.4$, and $r/R = 0.6$, the silica cap having the thickness of 12 nm (Fig. 1.17). According to our calculations, the largest linear dimension of the measured nanoparticle (including the shell) was about 200 nm, in good agreement with an estimate [65]. Note that the right shoulder of the main peak is due to actual

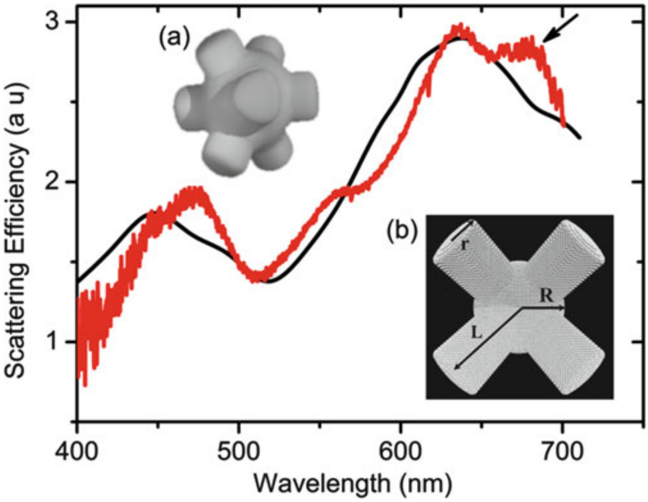


Fig. 1.17 Extinction spectra. *Red jagged curve*: data [65] for a particle in silica bead, *black*: theoretical fit with $a_{\text{eff}} = 70.5$ nm, $L/R = 1.4$ and $r/R = 0.6$, and the thickness of the bead 12 nm. *Arrow* indicates a shoulder reproducible with the arms having triangular cross section. *Insets*: (**a**) SEM image of a Ag octopod from [65]. (**b**) A view of a typical octopod along the fourth-order symmetry axis, C_4, where R, r, and L are the geometric parameters

cross section of the arms being close to a *triangular* shape rather than the round one used in Fig. 1.17b.

The way optical spectra evolve in going from nanosphere to nanostars, we present an extinction efficiency for nanoparticles with L/R ratio ranging from 1 (a sphere) to 2.4, for the case when the direction of the incident beam is directed along the most symmetrical axis C_4, Fig. 1.18. All nanoparticles have the same volume as a 81.4 nm sphere. The sphere exhibits two well-known peaks [56], the one at 375 nm corresponding to a quadrupole plasmonic resonance, whereas the second at 460 nm is dipolar in nature. The quadrupole plasmonic resonance starts splitting into two when arms appear, as it has due to reduction of the initial spherical symmetry O_3 to cubic O_h. One of these two new peaks (marked as 1) becomes noticeably lower and slightly blue shifts. Another one (indicated as 2), on the contrary, increases in height and red shifts. The dipole resonance at 460 nm does not split under the reduction of symmetry, but quickly red shifts with L/R (the corresponding maximum is marked as 3). This maximum gets progressively higher and at some moment surpasses the height of the second peak.

Interestingly, the red-most quadrupole resonance 2 is actually an *absorption* peak, which does not coincide with the scattering resonance. Similarly, the dipole peak 3 comes, in fact, from the *scattering* component, not from an absorption contribution (similar to the dipole peak for the pure *sphere* in Fig. 1.18). The fact that the absorption and scattering cross sections do not peak simultaneously means that both the SPRs (2 and 3) are *virtual*, i.e., correspond to the complex frequencies [69].

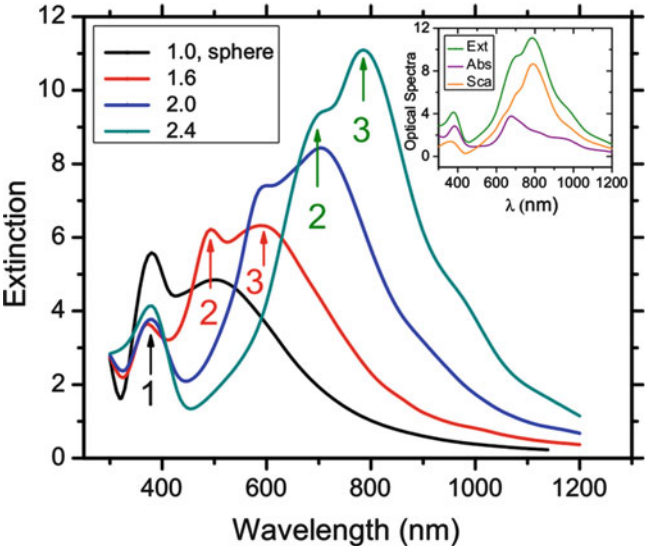


Fig. 1.18 Extinction spectra of sphere and multiarm nanoparticles, all with the same effective radius 81.4 nm. The ratio of L/R is, *from left to the right*: 1 (sphere), 1.6, 2.0, and 2.4. All the nanostars have the same ratio $r/R = 0.5$. The *inset* shows decomposition of the extinction efficiency into absorption and scattering contributions for $L/R = 2.4$ as a function of the wavelength λ

Due to high cubic symmetry, the optical properties of the nanostars are expected to be only weakly dependent on the direction of the incident beam **k**. To prove this, we compared the extinction spectra for three different **k** vectors directed along the symmetry axes C_2, C_3, and C_4. When averaged over the two mutually orthogonal directions of polarizations, the extinction spectra (not shown here) exhibit remarkable closeness to each other. This means that *any* beam can excite all the plasmonic modes in the system provided that the light is elliptically or circularly polarized. If the light is linearly polarized, there are only few k directions for which the spectrum is noticeably different for different directions of polarizations. These include the case when **k** is parallel to the second-order axis, C_2 or [110], with the polarizations along the [1−10] and [001] directions.

Now, we are going to discuss the electromagnetic modes associated with the SPRs 1, 2, and 3 for the case $L/R = 2.4$ in Fig. 1.18 in more detail. Each mode can be characterized by a dipole distribution, which oscillates in time; such distributions corresponding to the initial moment of time are shown in Fig. 1.19 (panels 1a–3a). It is clearly seen that the resonance 1 is a quadrupole in nature, because here approximately half of the electron cloud moves mainly up (red arrows), perpendicular to the axis of arms, while another half moves mainly down (blue arrows), along the arms. This resonance can be considered as a *bonding hybridization* [63] between the initial sphere quadrupole resonance and the arms' dipolar modes polarized both *parallel* and *perpendicular* to the cylindrical axes.

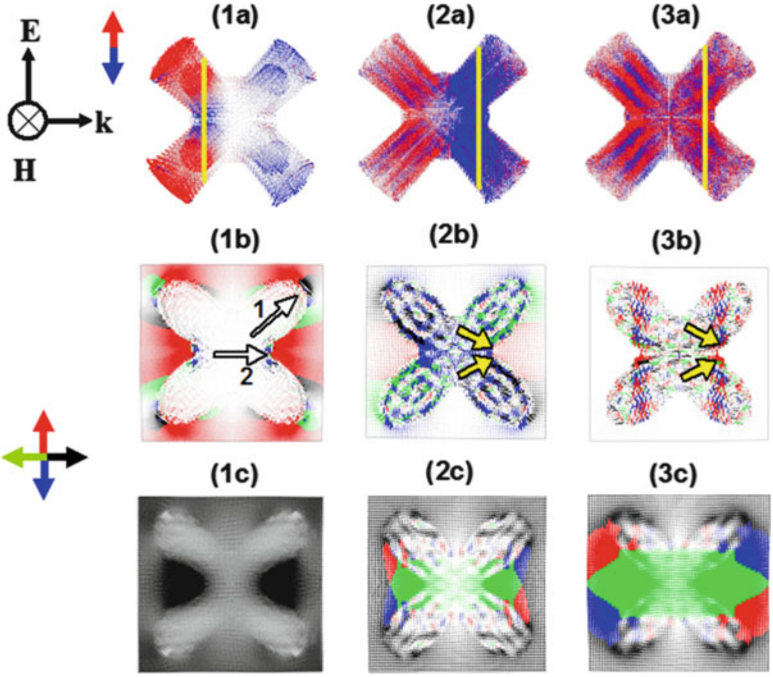


Fig. 1.19 (**a**) 3D dipole distributions at time $t = 0$ for the SPRs 1, 2, and 3. The *red* and *blue arrows* show local dipoles directed along or against the external electric field. The hot spots lie in the planes marked by the *yellow lines*. (**b**) 2D electric field distributions on the "hot spot" planes. Shown in *red*, *blue*, *green*, and *black* are the vectors **E** directed predominantly *up*, *down*, *left*, and *right*, respectively. The length of the *arrows* is proportional to |**E**|. The *big arrows* mark the hot spots. (**c**) Same as in (**b**), but for the magnetic field

For the *second* plasmon, the resonant currents flow up and down along the arms, but in *opposite* directions for the first four arms (hit by the beam first) compared to the rest four arms (reached by the beam later). Under the symmetry operations of the point group O_h, the current field transforms as a row of the three-dimensional irreducible unitary representation Γ'_{25} (T_{2g}) [70]. Therefore, this is a threefold degenerate *quadrupole resonance*. And finally, in the SPR 3, the local dipoles oscillate in a similar way in all eight arms. This is, of course, a threefold degenerate dipole resonance with Γ_{15} (T_{1u}) [70] symmetry (transforming as components of a vector).

It is convenient to present hot spots or maxima $|\mathbf{E}|^2$ induced by the resonances on the planes perpendicular to the direction of **k**. The positions of such a plane can be characterized by their dimensionless shortest separation x (measured in units of R) from the center of the nanostar. We attribute two signs to x: $x < 0$, if the light beam hits the plane earlier in time than the center, and $x > 0$, if later. The planes contained maximal $|\mathbf{E}|^2$ for the critical frequencies 1, 2, and 3 are defined by $x = -0.80$, 0.85, and 0.80, respectively. They are marked by the yellow lines in Fig. 1.19 (panel a).

For the SPR 1, the field is localized on the surface leading to eight hot spots of two different groups (4 spots in each), Fig. 1.19 (panel 1b). The spots of first group are located on the arms not far from their ends and associated with the local dipolar charge distributions directed *perpendicular* to the arms. The spots of second group reside between the arms near their confluence; they are induced by the electric current flowing between the upper and lower arms. The near-field enhancement factor associated with the both groups of spots is comparatively modest ($|\mathbf{E}|/|\mathbf{E}_0| \approx 10$).

In the case of SPR 2, the field maxima are obtained near the junctions of the arms located *transversely* to the polarization of the incident light, Fig. 1.19 (panel 2b). At these points, the field is directed tangentially to the surface and opposite to the incident field producing the enhancement factor of $|\mathbf{E}|/|\mathbf{E}_0|$ on the order of 40. Accidentally, the positions of hot spots associated with the dipolar resonance 3 (panel 3b) are very close to those of the resonance 2. However, in passing from the SPR 2 to 3, the character of the electric field changes drastically. Now, in the hot spot region, the field tends to be *parallel* to the polarization of the incident light. Besides, it becomes more complicated in the core region displaying an involved pattern. This resonance produces the *largest* enhancement factor of the field enhancement of about 50.

To understand the origin of electric hot spots, it is necessary to consider the distribution of magnetic fields on the same *x*-planes containing electric hot spots (Fig. 1.19, panel c). As is seen from the figure, the maxima in $|\mathbf{H}|^2$ practically coincide with those in $|\mathbf{E}|^2$, with the only exception of SPR 1, where the spots of the first group do not have their magnetic analogs (compare the panels 1a and 1b). This means that the hot spots of the second group along with all the other spots (corresponding to the SPRs 2 and 3) are actually magneto*inductive* in nature, stemming from the oscillating virtual current loops where the current inside the particle is shunted by the fields outside of the particle. This conclusion is not surprising, because the linear size of the particles is 5–7 times *bigger* than the wavelength inside them (for the wavelengths in vacuum λ in the range of 600–800 nm). But in such a case, as pointed by Landau and Lifshitz [26], the scattering properties of the particles are not defined by the excited *electrical* multipoles alone, and the induced eddy currents (*magnetic* dipoles) become equally important. It is interesting that in the region of magnetic hot spots, the field $\mathbf{H}(\mathbf{r})$ is directed along (SPR 1) or against (SPRs 2 and 3) that of incoming light. The latter would correspond to an effective negative permeability at optical frequencies. Besides, it may be possible to directly visualize those hot *magnetic spots* by trapping small magnetic particles in their vicinity.

It is remarkable that for relatively large *L*/*R*, the SPRs 2 and 3 are separated practically by the same wavelength distance (~100 nm). This trend tends to be preserved even if the geometry of the arms is changed, although the relative strength of the plasmons is strongly sensitive both to the arm length and to its radius. Thus, with increasing *r*, the peak 2 in the extinction curve becomes higher and sharper, while the peak 3 is getting lower. By playing only with *L* and *r,* one may engineer a two-peak resonant structure in the wide region of working wave lengths between 650 and 900 nm.

The optical properties of silver nanostars with cubic symmetry offer new interesting possibilities in comparison with the particles of other shapes. First, due to their high symmetry, their plasmonic modes can be excited practically by any light beam, independently of its polarization and direction of propagation. Therefore, the mutual orientation of the target and light becomes much less important than in the case of nonsymmetric stars. Second, it is natural to expect that when a molecule is deposited on a nanostar, it may stick somewhere between the arms near their junction. This means that the molecule will find a favorable local environment (hot spots) characterized by strongly enhanced electric fields. And third, the Raman scattering cross section scales roughly as $|\mathbf{E}(\omega)|^2$ $|\mathbf{E}(\omega \pm \Omega)|^2$, where Ω is the vibrational frequency of the molecule [48]. So, it is important to have a large electric field not only at excited frequency ω but also at Raman shifted scattering frequencies $\omega_R = \omega \pm \Omega$, and the resonances with close frequencies can provide such a realization. In this context, it is very important that the hot spots induced by these resonances are not separated in space but practically coincide.

1.10 Efficient SERS Substrate with Collapsed Gold Nanofingers

Perhaps most interesting for enhanced spectroscopy are various assemblies of closely spaced nanoparticles. The assemblies with narrow gaps between constituent particles may support very intense local fields, which are very advantageous for SERS, etc. Precise control of both the shape and geometrical symmetry of the assemblies has been an active focus of many research efforts. For example, by using either top-down or bottom-up approaches, nanoparticle dimers (see e.g., [52, 72–76]) and trimers [77] have been studied rather extensively. However, each of the two approaches has certain limitations. The top-down approach allows one to introduce symmetry and gap size control at certain scale, but sub-5 nm gap sizes have been elusive. On the other hand, the bottom-up approach can achieve small gap sizes, but it is difficult to achieve highly uniform structures of arbitrary symmetries across a large area, such as tetramers, pentamers, or heptamers, for systematic study. Yet complex structures such as these have extremely interesting photonic properties; for example, it has been recently shown that when nanoparticles or nanopillars are assembled in a heptamer, one can observe Fano-like resonance [76–79]. It is therefore extremely interesting to study the plasmonic properties of these more complicated assemblies and also examine their performance for various applications, such as SERS sensing. Despite a significant amount of progress in the exploration of dimers [36, 76, 80, 81], few extensive studies have been performed on the higher order assemblies [77] because of the difficulty in fabrication.

We have recently performed systematic studies of the plasmonic properties of regular polygons, including 2-mer (digon), 3-mer (trigon), 4-mer (tetragon), 5-mer

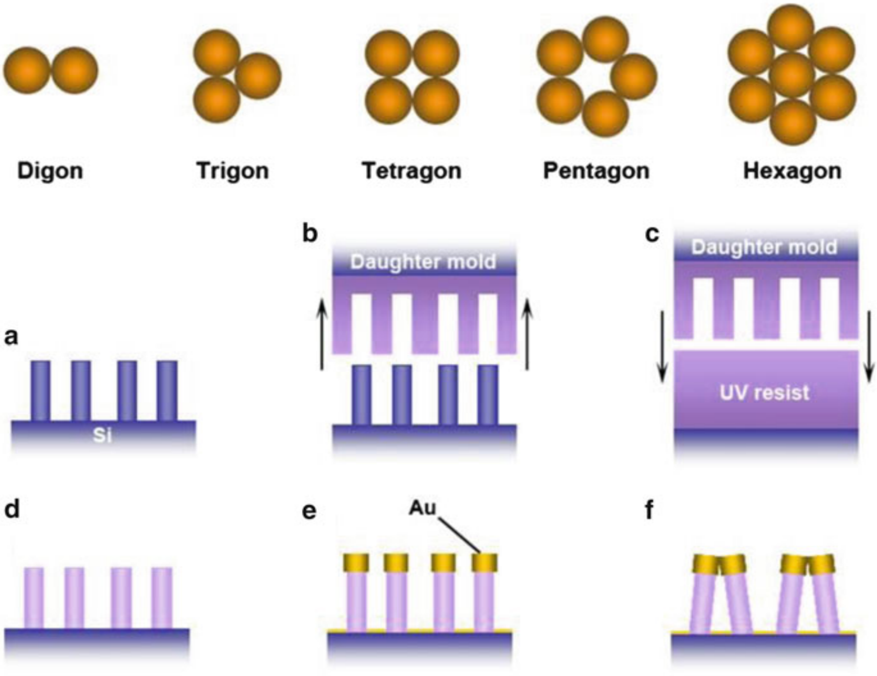


Fig. 1.20 Schematic drawings of the metal nanoparticle cap assemblies formed after *closure* of supporting their nanofingers, from dimer (digon) through to hexagon with seven nanoparticles (*top panel*). (**a**)–(**f**) Fabrication procedure for the nanofingers: (**a**) fabrication of the nanofinger silicon mold using E-beam lithography, (**b**) making of the daughter mold using nanoimprinting, (**c**), (**d**) fabrication of the polymer nanofingers from the polymer daughter mold using nanoimprinting. (**e**) E-beam deposition of 80 nm Au onto the nanofingers. (**f**) soaking and drying of the solvent to induce the closure of the nanofingers

(pentagon), and 7-mer (hexagon arrangement with the seventh particle in the center), and the SERS signal that arises from them [82], see Fig. 1.20 (top panel). These structures have been fabricated using nanoimprinting technique described in [83], Fig. 1.20a–f. The crux of the matter that helped obtaining extraordinary SERS signal enhancement was the closure of nanofinger structures after the fingers have been wet with solvent that dried out and capillary forces facilitated the closure of the assembly, Fig. 1.20f.

All those structures have been simulated by DDA and FDTD methods that produced similar results when applied to the same structures and helped in gaining more insight into the observed enhanced performance. The gold caps have puck-like or close to a hemispherical shape, which we used in simulations. The small separation ensures a strong coupling between the neighboring metal nanospheres so that the electromagnetic near-field is greatly enhanced [36, 53]. The symmetries of the metal nanostructures lead to interesting plasmonic modes, in analogy with molecular-orbital hybridization in conjugated systems, first reported in [82].

Here, polygon-shaped gold nanostructures with various symmetries were formed on top of predefined flexible polymer fingers, which then self-assemble with the aid of microcapillary forces.

In greater detail, the experimental procedure for the fabrication of the gold-coated nanofingers is illustrated by Fig. 1.20 [82]. The UV-curable NIL with double-layer spin-on resists were utilized for all NIL steps, performed with a custom-designed nanoimprint machine [80]. A polymeric reverse-tone mold was used to duplicate the Si pillars into the final polymer nanofingers shown in Fig. 1.20d, with the procedure similar to that previously reported [81, 84]. Finally, Au with nominal thickness of 50–80 nm was deposited on the sample by E-beam evaporation at normal incidence to form the Au nanoparticles on the tips of the polymer fingers, Fig. 1.2e. After the arrays have been exposed to solvent and air dried, the fingers closed together in the designated symmetry as shown in Fig. 1.2f. In all geometries, the fingers were about 520 nm in height before coating by Au. Each assembly separated from other group by 200 nm, with the periods of the 2-mer, 3-mer, and 4-mer groups 500 nm, while the periods of the 5-mer and the 7-mer groups were 700 nm. The *capillary force* is believed to be the main driving force that leads to the coalescence of the nanofingers; a similar phenomenon was observed in high aspect ratio microscale structures [85–87]. The final geometries are clear from the top and side view SEM images in Fig. 1.21.

In order to study the symmetry dependence of the nanofinger assemblies, we chose a model molecule, *trans*-1,2-bis(4-pyridyl)-ethylene (BPE), for SERS studies. The integration time was 1 s for the entire spectrum. The Raman spectra of BPE excited at 785 nm when using digon, trigon, tetragon, pentagon, hexagon, and dot arrays as the control for the SERS substrates are shown in Fig. 1.22a. The nanofingers on the periodic dot array control sample formed random assemblies after the closing process. All the substrates generated qualitatively the same sets of Raman peaks with similar peak intensity ratios from the test molecule, BPE, suggesting that there is no significant difference in terms of the enhancement mechanism from the various symmetries finger assemblies.

When comparing the absolute intensity of the Raman peaks from different symmetries, it is clear that *pentamers* outperformed the rest of the symmetry designs. Note the intensity of the Raman signal was normalized against the number of pillars per unit area; hence, the figure shows the intensity/finger in the given designs. Figure 1.22b shows comparison of the normalized intensity of the Raman signal at 1,600 cm^{-1} measured at three different laser wavelengths. As in the case of using the laser at 785 nm, the pentagon outperformed the rest of the geometries at 633 nm and 1,064 nm excitation wavelengths as well. One can calculate the enhancement factor to be ~10^{11} for the nanofinger pentagon, a further improvement of about a factor of 6 when compared to the random nanofinger assemblies (dot array sample in Fig. 1.22b) [82].

In order to analyze the underlying physics of the observed enhancement, simulations were performed using the discrete-dipole approximation (DDA), which is applicable to any arbitrary particle shape and configuration of particles. The dielectric constants are described using the empirical data for Au for different

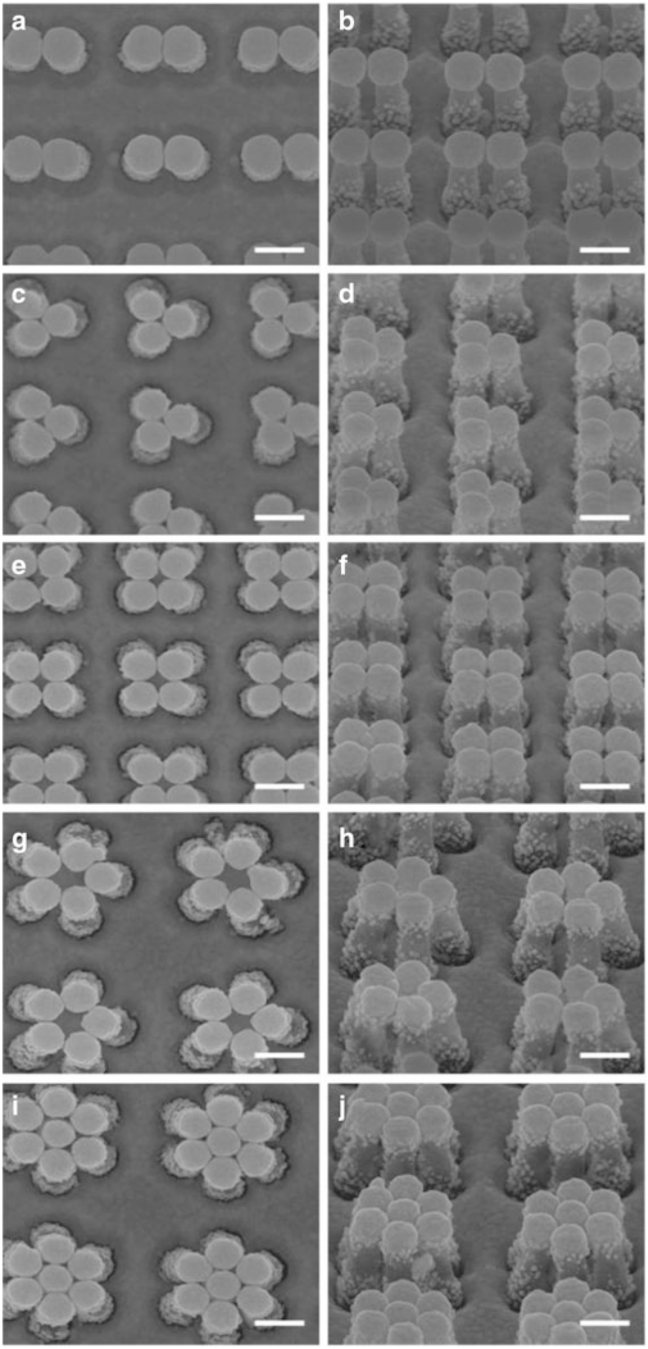


Fig. 1.21 SEM image showing the top view and the side view (45° tilt from the surface normal) of the closed nanofinger assemblies for (**a**, **b**) digon, (**c**, **d**) trigon, (**e**, **f**) tetragon, (**g**, **h**) pentagon, and (**i**, **j**) hexagon-shaped 7-mer. *Scale bars* in the SEM images are 200 nm

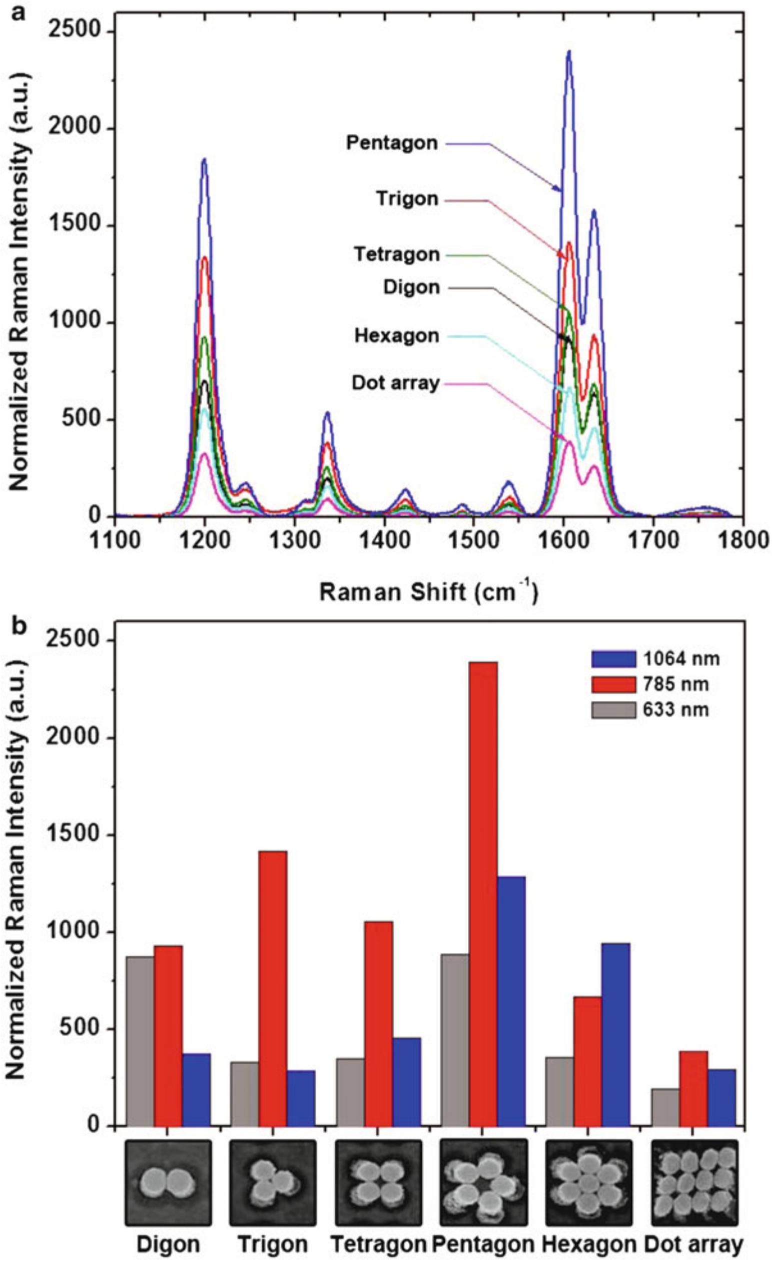


Fig. 1.22 (**a**) Normalized Raman spectra of BPE excited at 785 nm for the digon, trigon, tetragon, pentagon, hexagon, and the dense array of nanofingers. (**b**) Comparison of the normalized intensity of the Raman signal at 1,600 cm^{-1}measured for 633 nm, 785 nm, and 1,064 nm incident radiation

wavelengths [31]. The distributions of electric field $|\mathbf{E}(\mathbf{r})|^2$ for the various symmetries corresponding to 785 nm along with the representative SEM images of individual polygons are shown in Fig. 1.23. The "hot spots" or maxima in $|\mathbf{E}(\mathbf{r})|^2$ are located precisely in the gaps between the spheres and characterized by near-field enhancement factors E/E_0 ranging from 60 (7-mers) to 180 (2-mers). The detailed instantaneous distributions of electric field inside the particles are rather complicated and characterized by a fine stripe intensity "domain" structure. This is due to the fact that the linear size of the particles is ~10 times bigger than the effective wavelength inside them. The larger-scale structure represents the amplitude of the multiparticle plasmon resonances that can be excited by radiation incident normal to the plane of the polygon.

Nordlander and co-workers have analyzed the group theoretic properties of the plasmons for the trimer through the heptamer structures [75, 88, 89]. We can qualitatively understand the experimental results and the DDA field intensity calculations from the symmetry-adapted basis functions, based on two in-plane dipoles for each atom in the structure, and their irreducible representations. Although there are twice as many basis functions as atoms for each structure, only a subset of those for each structure, belonging to a particular irreducible representation for the space group of the polygon, have the correct symmetry to couple to the in-plane polarization vector of the incident radiation. Within the set of basis functions that are dipole active, some of those may have nodes that pass between a pair or pairs of atoms in the structure, and thus those potential hot spots are symmetry forbidden to "light up," cf. Fig. 1.23e for the heptamer—all of the potential hot spots that exist between the central particle and the six surrounding particles are dark and thus do not contribute to Raman enhancement of trapped molecules. Furthermore, because of the high symmetry of the heptamer and the requirement that the basis functions that couple to dipolar excitation must be odd, the extent of the in-phase overlap of the dipoles between the particles on the ring of the heptamer is limited. This gives rise to the relatively weak calculated field enhancements, Fig. 1.23e. In contrast, for the lower symmetry pentagon structure, the basis functions of the dipole-allowed irreducible representation have significant overlap between nearly every pair of atoms in the structure, which leads to the extremely strong plasmon excitation for this structure compared to the other polygons.

The group theory alone cannot be used to predict quantitative field enhancement differences between the various polygons. That requires a detailed understanding of the material properties and structural details of the polygons and the wavelength of the incident light. For the heptamer, there is the additional issue that structures with an atom inside a polygon may exhibit a Fano resonance [77], which adds a level of complexity to the spectral response of the structure. In particular, there can be a strong antiresonance between alternative excitation paths that will dampen the plasmon resonance of the heptamer and thus make it less attractive for SERS.

Here, we described a simple and scalable method to produce Au nanoparticles on flexible nanofingers that can be closed by capillary forces producing extremely strong Raman signal. This high precision to assemble nanostructures opens a new

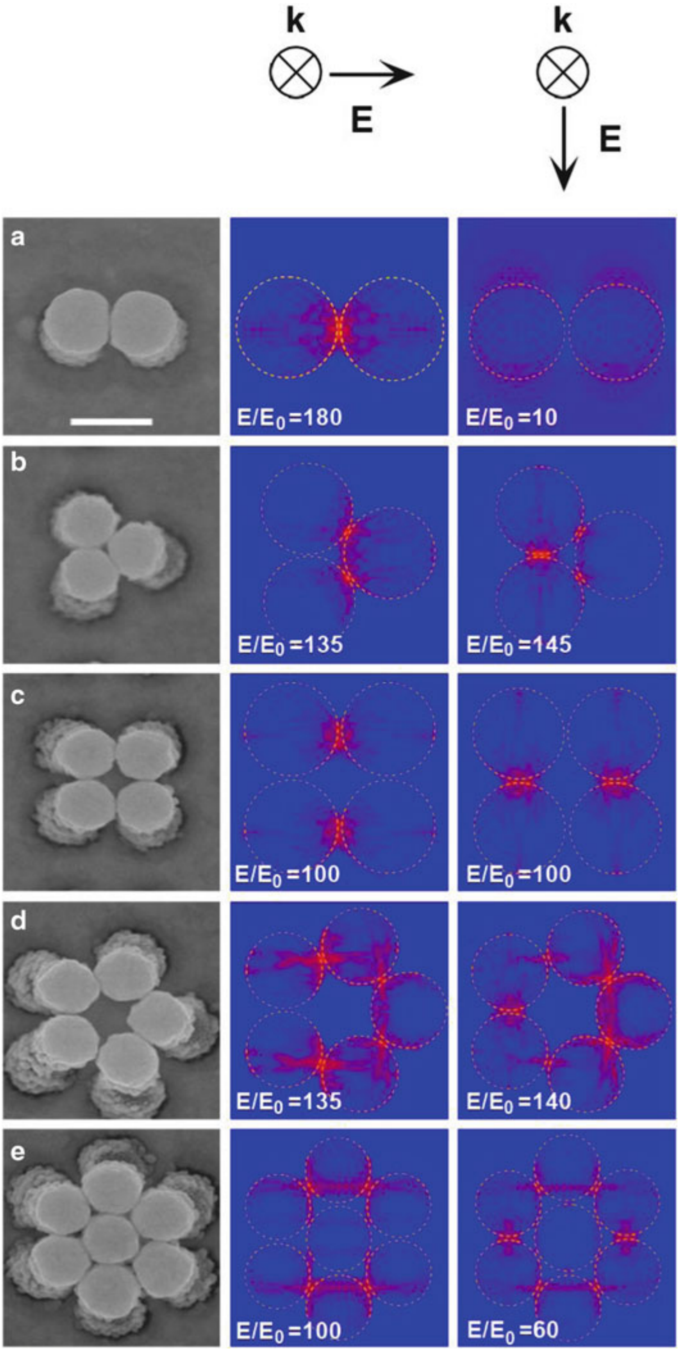


Fig. 1.23 (**a**)–(**e**) Distributions of electric field intensity $|\mathbf{E}|^2$ at 785 nm in the central plane of all the polygon shaped assemblies shown in Figure 1.20 for the two different in-plane electric field polarizations illustrated on the *top panel*. *Scale bar* in SEM is 200 nm

path for the design and fabrication of arbitrary geometries of nanostructures. Interestingly, the polygon symmetry determines the number and the strength of the potential hot spots between Au particles that actually light up and thus contribute to SERS. The elementary electromagnetic modes (plasmons) in pentagon, in particular, have symmetry (with dipole component) allowing them to couple well to external field and provide significant in-phase field overlap between adjacent particles. This suggests a means for further engineering multiparticle plasmon resonances by exploring various broken symmetries.

1.11 Conclusions

We have described the hallmark properties of NIM, from negative refraction to superlensing, and concentrated on a particular NIM system, a so-called fishnet metallic–dielectric structure, where *both* the dielectric permittivity and magnetic permeability can become negative. Since one of the goals of this chapter is to describe possibilities for fast optical modulation with metamaterials, hence we have looked into the origin of the fast carrier response in a fishnet structure. We speculate that fast surface relaxation of photocarriers in thin α-Si spacer layer may indeed proceed within a few picoseconds while other processes (interband transitions and deep level trapping in semiconducting spacer layer) may involve longer time scales, on the order of a few tens of picoseconds. More experimentation is desirable to gain more insight into the effect of gain and mechanism of ultrafast photocarrier relaxation in those metamaterials.

While thinking about applications of metamaterials in (nano)photonics, one may identify two problems (1) to make the form-factor for optical components compatible with that of CMOS circuitry and (2) to use the fabrication steps compatible with Si processing, enabling the eventual monolithic integration of optical components that are based on metamaterials. This effort is to be viewed as a part of silicon photonics that would ultimately need CMOS-compatible Si-based light sources, high-index waveguides, high-speed modulators, and photodetectors. One should bear in mind that metal-based plasmonic devices necessarily exhibit very large losses in the infrared and visible range, therefore, the whole chips outfitted with plasmonic waveguides, etc. would not be practical, and plasmonic lasers (spasers) are unlikely [90], but one can think of using isolated plasmonic elements, where loss does not matter. The choice of materials for plasmonic applications at visible wavelengths is also pretty much limited to Silver and Gold [91].

As reviewed above, plasmonic structures and metamaterials provide for very fast optoelectronic devices but suffer from losses that limit their applications [94]. The field where the large plasmonic losses in the infrared and visible ranges are not important, while a huge dielectric constant contrast is (and the accompanying very large local field enhancement) belongs in sensors [95, 97]. Recent work brought about tremendous advances in sensors that use local field enhancement due to plasmon resonances in neighboring metallic structures, which is especially true of

SERS growing as the fourth power of the local field [96, 97]. Manipulating the nanostructures in a controlled way, like the above nanofingers with gold caps, it is possible to trap molecules in narrow gaps between the metallic caps, thus facilitating ultimate sensitivity to this technique. No doubt, one should expect more breakthroughs in controllable molecule trapping and analysis in near future to pave the way for a wide usage of SERS and other plasmonics techniques in sensing.

References

1. L.I. Mandelshtam, *Lectures in Optics, Relativity, and Quantum Mechanics* (Nauka, Moscow, 1972), p. 389
2. R.A. Silin, V.P. Sazonov, *Delay Systems* (Sov. Radio, Moscow, 1966), Ch. 8
3. R.A. Silin, Phys. Uspekhi **175**, 562 (2006)
4. V.Veselago, Usp. Fiz. Nauk **92**, 517 (1967) [Sov. Phys. Usp. **10**, 509 (1968)]
5. J.B. Pendry, Phys. Rev. Lett. **85**, 3966 (2000)
6. N. Fang, H. Lee, C. Sun, X. Zhang, Science **308**, 534 (2005)
7. D.O.S. Melville, R.J. Blaikie, C.R. Wolf, Appl. Phys. Lett. **84**, 4403 (2004)
8. U. Leonhardt, Science **312**, 1777 (2006)
9. U. Leonhardt, T. Tyc, Science **323**, 110 (2009)
10. J.B. Pendry, D. Schurig, D.R. Smith, Science **312**, 1780 (2006)
11. D. Schurig, J.B. Pendry, D.R. Smith, Opt. Express **14**, 9794 (2006)
12. D. Schurig, J.J. Mock, B.J. Justice, S.A. Cummer, J.B. Pendry, A.F. Starr, D.R. Smith, Science **314**, 977 (2006)
13. T. Ergin, N. Stenger, P. Brenner, J.B. Pendry, M. Wegener, Science **328**, 337 (2010)
14. A.V. Rogacheva, V.A. Fedotov, A.S. Schwanecke, N.I. Zheludev, Phys. Rev. Lett. **97**, 177401 (2006)
15. I.I. Smolyaninov, New J. Phys. **5**, 147 (2003)
16. D.A. Genov, S. Zhang, X. Zhang, Nat. Phys. **5**, 687 (2009)
17. E. Kim, W. Wu, E. Ponizovskaya, Z. Yu, A.M. Bratkovsky, S.-Y. Wang, R.S. Williams, Y.R. Shen, Appl. Phys. Lett. **91**, 173105 (2007)
18. J.A. Schuller, E.S. Barnard, W. Cai, Y.C. Jun, J.S. White, M.L. Brongersma, Nat. Mater. **9**, 193 (2010)
19. R. Pratibha, K. Park, I.I. Smalyukh, W. Park, Opt. Express **17**, 19459 (2009)
20. A.B. Golovin, O.D. Lavrentovich, Appl. Phys. Lett. **95**, 254104 (2009)
21. Z.L. Sámson, K.F. MacDonald, F. De Angelis, B. Gholipour, K. Knight, C.C. Huang, E. Di Fabrizio, D.W. Hewak, N.I. Zheludev, Appl. Phys. Lett. **96**, 143105 (2010)
22. S. Zhang, W. Fan, K.J. Malloy, S.R.J. Brueck, N.C. Panoiu, R.M. Osgood, Opt. Express **13**, 4922 (2005)
23. V.M. Shalaev, W. Cai, U.K. Chettiar, H.-K. Yuan, A.N. Sarychev, V.P. Drachev, A.V. Kildishev, Opt. Lett. **30**, 3356 (2005)
24. G. Dolling, M. Wegener, C.M. Soukoulis, S. Linden, Opt. Lett. **32**, 53 (2007)
25. V. Pafomov, Zh. Eksp. Teor. Fiz. **36**, 1853 (1959)
26. L.D. Landau, E.M. Lifshitz, *Electrodynamics of Continuous Media* (Pergamon, New York, 1993)
27. M. Notomi, Theory of light propagation in strongly modulated photonic crystals: Refraction-like behavior in the vicinity of the photonic band gap. Phys. Rev. B **62**, 10696–10705 (2000)
28. F.D.M. Haldane, *Electromagnetic Surface Modes at Interfaces with Negative Refractive Index Make a Not-Quite-Perfect Lens*. cond-mat/0206420v3 (2002)

29. A.M. Bratkovsky, A. Cano, A.P. Levanyuk, Appl. Phys. Lett. **87**, 103507 (2005)
30. N. Garcia, M. Nieto-Vesperinas, Phys. Rev. Lett. **88**, 207403 (2002); **90**, 229903 (E) (2003)
31. P.B. Johnson, R.W. Christy, Phys. Rev. B. **6**, 4370 (1972)
32. E.D. Palik, *Handbook of Optical Constants of Solids* (Academic Press, New York, 1998)
33. W. Wu, E. Ponizovskaya, E. Kim, D. Chou, A. Bratkovsky, Z. Yu, Q. Xia, X. Li, Y.R. Shen, S.Y. Wang, R.S. Williams, Appl. Phys A **87**, 143 (2007)
34. A. Taflove, S.C. Hagness, *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, 3rd edn. (Artech House, Boston, 2005)
35. E.V. Ponizovskaya, A.M. Bratkovsky, Appl. Phys. A **95**, 1137 (2009)
36. J.P. Camden, J.A. Dieringer, Y. Wang, D.J. Masiello, L.D. Marks, G.C. Schatz, R.P. Van Duyne, J. Am. Chem. Soc. **30**, 12617 (2008)
37. D.R. Smith, S. Schultz, P. Markos, C.M. Soukoulis, Phys. Rev. B **65**, 195104 (2002)
38. H. Bergner, V. Brueckner, B. Schroeder, Sov. J. Quant. Electron. **13**, 736 (1983)
39. A. Esser, K. Seibert, H. Kurz, G.N. Parsons, C. Wang, B.N. Davidson, G. Lucovsky, R.J. Nemanich, Phys. Rev. B **41**, 2879 (1990)
40. L.A. Coldren, S.W. Corzine, *Diode Lasers and Photonic Integrated Circuits* (Wiley, New York, 1995)
41. D.J. Cho, W. Wu, E. Ponizovskaya, P. Chaturvedi, A.M. Bratkovsky, S.-Y. Wang, X. Zhang, F. Wang, Y.R. Shen, Opt. Express **17**, 17652 (2009)
42. S. Adachi, Phys. Rev. B **38**, 12966 (1988)
43. See http://www.ioffe.ru/SVA/NSM/nk/GermaniumCompounds/Gif/ge.gif.
44. S.A. Matsumoto, R.I. Joseph, R.M. Sova, M.E. Thomas, Proc. SPIE **5620**, 228 (2004)
45. M. Fleischmann, P.J. Hendra, A.J. McQuillan, Chem. Phys. Lett. **66**, 163 (1974)
46. D.L. Jeanmaire, R.P. Van Duyne, J. Electroanal. Chem. **84**, 1 (1977)
47. M.G. Albrecht, J.A. Creighton, J. Am. Chem. Soc. **99**, 5215 (1977)
48. E. Le Rue, P.G. Etchegoin, *Principles of Surface Enhanced Raman Specroscopy and Related Plasmonic Effects* (Elsevier, Amsterdam, 2009)
49. M. Moskovits, J. Raman Spectrosc. **36**, 485 (2005)
50. M. Moskovits, Rev. Mod. Phys. **57**, 783 (1985)
51. D.A. Long, *The Raman Effect* (Wiley, New York, 2001)
52. C. Sonnichsen, B. Reinhard, J. Liphardt, A. Alivisatos, Nat. Biotechnol. **23**, 741 (2005)
53. J. Jiang, K. Bosnick, M. Maillard, L. Brus, J. Phys. Chem. B **107**, 9964 (2003)
54. J.P. Kottmann, O.J.F. Martin, D.R. Smith, S. Schultz, New J. Phys. **2**, 271 (2000)
55. Y. Sun, Y. Xia, Science **298**, 2176 (2002)
56. K.L. Kelly, E. Coronado, L.L. Zhao, G.C. Schatz, J. Phys. Chem. B **107**, 668 (2003)
57. B.J. Wiley, S.H. Im, Z.-Y. Li, J. McLellan, A. Siekkinen, Y. Xia, J. Phys. Chem. B **110**, 15666 (2006)
58. Y. Chen, C. Wang, Z. Ma, Z. Su, Nanotechnology **18**, 325602 (2007)
59. J.Z. Zhang, C. Noguez, Plasmonics **3**, 127 (2008)
60. I.O. Sosa, C. Noguez, R.G. Barrera, J. Phys. Chem. B **107**, 6269 (2003)
61. B. Wiley, Y. Sun, Y. Xia, Acc. Chem. Res. **40**, 1067 (2007)
62. C.L. Nehl, H.W. Liao, J.H. Hafner, Nano Lett. **6**, 683 (2006)
63. A. Tao, P. Sinsermsuksakul, P. Yang, Angew. Chem. Int. Ed. **45**, 4597 (2006)
64. F. Hao, C.L. Nehl, J.H. Hafner, P. Nordlander, Nano Lett. **7**, 729 (2007)
65. L. Rodríguez-Lorenzo, R. Álvarez-Puebla, I. Pastoriza-Santos, S. Mazzucco, O. Stéphan, M. Kociak, L.M. Liz-Marzán, F.J. García de Abajo, J. Am. Chem. Soc. **131**, 4616 (2009)
66. M.J. Mulvihill, X. Yi Ling, J. Henzie, P. Yang, J. Am. Chem. Soc. **132**, 268 (2010)
67. E.M. Purcell, C.R. Pennypacker, Astrophys. J. **186**, 705 (1973)
68. B.T. Draine, P. J. Flatau, J. Opt. Soc. Am. A **11**, 1491 (1994); **25**, 2693 (2008); arXiv:1202.3424v2 [physics.comp-ph] (2012)
69. A. Wokaun, J.P. Gordon, P.F. Liao, Phys. Rev. Lett. **48**, 957 (1982)
70. C.F. Bohren, D.R. Huffman, *Absorption and Scattering of Light by Small Particles* (Wiley, New York, 1998)

71. See e.g., A. Nussbaum, *Applied Group Theory for Chemists, Physicists and Engineers* (Englewood Cliffs, NJ, 1971)
72. E. Hao, G. Schatz, J. Chem. Phys. **120**, 357 (2004)
73. P. Nordlander, C. Oubre, E. Prodan, K. Li, M.I. Stockman, Nano Lett. **4**, 899 (2004)
74. L. Brown, H. Sobhani, J. Lassiter, P. Nordlander, N. Halas, ACS Nano **4**, 819 (2010)
75. N. Grillet, D. Manchon, F. Bertorelle, C. Bonnet, M. Broyer, E. Cottancin, J. Lerme, M. Hillenkamp, M. Pellarin, ACS Nano **5**, 9450 (2011)
76. K. Höflich, M. Becker, G. Leuchs, S. Christiansen, Nanotechnology **23**, 185303 (2012)
77. D.W. Brandl, N.A. Mirin, P.J. Nordlander, Phys. Chem. B **110**, 12302 (2006)
78. J.A. Fan, C.H. Wu, K. Bao, J.M. Bao, R. Bardhan, N.J. Halas, V.N. Manoharan, P. Nordlander, G. Shvets, F. Capasso, Science **312**, 1135 (2010)
79. J.A. Fan, K. Bao, C. Wu, J. Bao, R. Bardhan, N.J. Halas, V.N. Manoharan, G. Shvets, P. Nordlander, F. Capasso, Nano Lett. **10**, 4680 (2010)
80. N. Verellen, Y. Sonnefraud, H. Sobhani, F. Hao, G.A.E. Vandenbosch, V.V. Moshchalkov, P. Van Dorpe, P. Nordlander, S. Maier, ACS Nano **4**, 1664 (2010)
81. M. Hentschel, M. Saliba, R. Vogelgesang, H. Giessen, A.P. Alivisatos, N. Liu, Nano Lett **10**, 2721 (2010)
82. W. Wu, G.Y. Jung, D.L. Olynick, J. Straznicky, Z. Li, X. Li, D.A.A. Ohlberg, Y. Chen, S.Y. Wang, J.A. Liddle, W.M. Tong, R.S. Williams, Appl. Phys. A. Mater. Sci. Process. **80**, 1173 (2005)
83. H.X. Ge, W. Wu, Z. Li, G.Y. Jung, D. Olynick, Y.F. Chen, J.A. Liddle, S.Y. Wang, R.S. Williams, Nano Lett. **5**, 179 (2005)
84. F.S. Ou, M. Hu, I. Naumov, A. Kim, W. Wu, A.M. Bratkovsky, X. Li, R.S. Williams, Z. Li, Nano Lett. **11**, 2538 (2011)
85. M. Hu, F.S. Ou, W. Wu, I. Naumov, X. Li, A.M. Bratkovsky, R.S. Williams, Z. Li, J. Am. Chem. Soc. **132**, 12820 (2010)
86. W. Wu, M. Hu, F.S. Ou, Z. Li, R.S. Williams, Nanotechnology **21**, 25502 (2010)
87. D. Chandra, S. Yang, Langmuir **25**, 10430 (2009)
88. M. Kotera, N. Ochiai, Microelectron. Eng. **78–79**, 515 (2005)
89. S.F. Chini, A. Amirfazli, Langmuir **26**, 13707 (2010)
90. N.A. Mirin, K. Bao, P. Nordlander, J. Phys. Chem. A **113**, 4028 (2009)
91. K. Bao, N.A. Mirin, P. Nordlander, Appl. Phys. A. Mater. Sci. Process **100**, 333 (2010)
92. J.B. Khurgin, G. Sun, Appl. Phys. Lett. **99**, 211106 (2011)
93. P. Tassin, T. Koschny, M. Kafesaki, C. Soukoulis, Nat. Phot. March 4, 2012 (online), DOI: 10.1038/NPHOTON.2012.27
94. A.M. Bratkovsky, Negative index materials with gain media for fast optical modulation. Proc. IEEE **9**, 1317 (2009)
95. M.L. Brongersma, J.W. Hartman, H.A. Atwater, Phys. Rev. B **62**, R16356 (2000)
96. J. Anker, W.P. Hall, O. Lyandres, N.C. Shah, J. Zhao, R.P. Van Duyne, Nat. Mater. **7**, 442 (2008)
97. S.A. Maier, *Plasmonics: Fundamentals and Applications* (Springer, Berlin, 2007)

Chapter 2
Recent Advances in Nanoplasmonics and Magnetoplasmonics

Maxim R. Shcherbakov, Tatyana V. Dolgova, and Andrey A. Fedyanin

Abstract Nanoplasmonics is a vastly developing area of modern photonics, which is capable of providing mankind with new routes to fast and miniature communication and technologies. With unprecedentedly high bandwidth supplied by photons and subwavelength dimensions supplied by electrons, surface plasmon is the next candidate for the everyday-life information unit. In this chapter we review recent advances in controlling the generation and propagation of surface plasmon polaritons in nanostructured materials as well as utilization of surface plasmons in order to obtain efficient control over optical signals.

2.1 Surface Electromagnetic Waves

Surface electromagnetic waves are defined as waves propagating along the interface between two media and existing in both of them [1, 2]. The dispersion law of the surface wave can be derived from Maxwell's equations by substituting the solution in the form of a localized wave (the *Oz* axis is perpendicular to the interface and directed towards the first medium, the *Ox* axis is at the interface, Fig. 2.1):

$$A = A_0 e^{s_{1,2}z} e^{i(\omega t - k_{spp}x)}, \tag{2.1}$$

where k_{spp} is the wave vector of the surface wave and $s_{1,\,2}$ are the extinction coefficients in the first and second media, respectively. If $\varepsilon_1(\omega)$ and $\varepsilon_2(\omega)$ are

M.R. Shcherbakov (✉) • T.V. Dolgova • A.A. Fedyanin
Faculty of Physics, Lomonosov Moscow State University, Moscow 119991, Russia
e-mail: shcherbakov@nanolab.phys.msu.ru; dolgova@nanolab.phys.msu.ru; fedyanin@nanolab.phys.msu.ru

A. Korkin and D.J. Lockwood (eds.), *Nanoscale Applications for Information and Energy Systems*, Nanostructure Science and Technology,
DOI 10.1007/978-1-4614-5016-0_2, © Springer Science+Business Media New York 2013

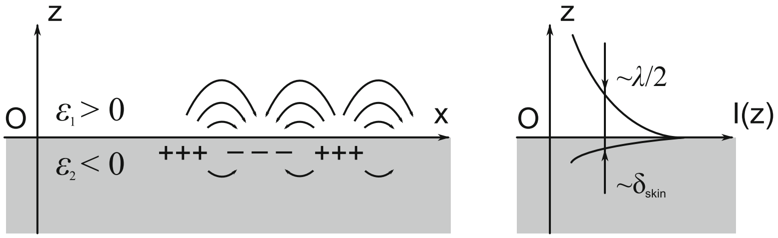


Fig. 2.1 *Left*: The interface between two media with the opposite signs of the real parts of the dielectric permittivity showing a schematic distribution of unbalanced charges and the electric field vector lines of the surface wave. *Right*: Intensity distribution and typical localization scales of the surface waves in the vicinity of the interface

complex dielectric constants of the first and second media, respectively, the dispersion relation takes the following form:

$$k_{spp}(\omega) = \frac{\omega}{c}\sqrt{\frac{\varepsilon_1(\omega)\varepsilon_2(\omega)}{\varepsilon_1(\omega)+\varepsilon_2(\omega)}}. \tag{2.2}$$

In the case of a metal surface in a dielectric environment, $k_{\rm spp}$ is an imaginary number and the surface wave is called surface plasmon polariton (SPP) since the unbalanced charges rise due to oscillations of free electrons of metal. In general $k_{\rm spp}$ can be written as:

$$k_{spp}(\omega) = k'(\omega) + ik''(\omega) = \frac{\omega}{c}\sqrt{\frac{\varepsilon_2'(\omega)\varepsilon_1}{\varepsilon_2'(\omega)+\varepsilon_1}} + \frac{i\omega}{c}\left(\frac{\varepsilon_2'(\omega)\varepsilon_1}{\varepsilon_2'(\omega)+\varepsilon_1}\right)^{3/2}\frac{\varepsilon_2''(\omega)}{2(\varepsilon_2'(\omega))^2}. \tag{2.3}$$

Figure 2.2 shows a typical dispersion law of SPPs at the metal–air interface ($\varepsilon_1 = 1$) in the form of the expression (2.2). Here the dielectric function of metal is approximated with the Drude model:

$$\varepsilon_2(\omega) = 1 - \frac{\omega_p^2}{\omega^2}. \tag{2.4}$$

It should be noted that this model describes well the optical response of metals in the infrared region. However, the model is not valid in the ultraviolet and visible spectral ranges due to the presence of pronounced resonances associated with the interband electron transitions. Besides, the ohmic losses in the metal are not taken into account.

The value of the imaginary part of the SPP wave vector indicates the energy losses due to its transfer from the oscillating electrons to the lattice. The mean free path of SPPs at a smooth metal surface is described by:

$$L = \frac{1}{2k''_{spp}}, \tag{2.5}$$

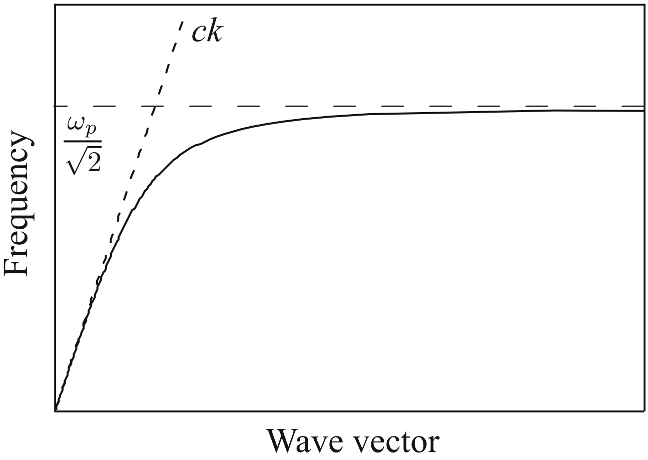


Fig. 2.2 The dispersion law of a surface plasmon polariton (*solid line*) and light in vacuum (*dotted line*)

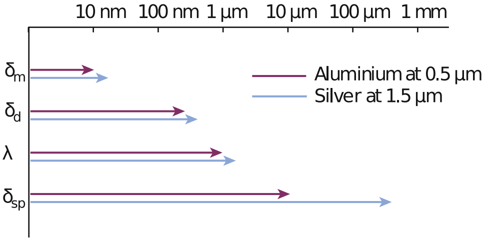


Fig. 2.3 Characteristic length scales of surface-plasmon-based photonics: the propagation length δ_{SP}, the decay length in the dielectric material δ_d, and the decay length in the metal δ_m [3]

where k_{spp} is the imaginary part of the SPP wave vector and is determined by the dielectric properties of the metal. For example, the free path of SPPs at the silver surface is approximately 1 mm for the telecommunication wavelength range ($\lambda \simeq 1.5$ µm). Figure 2.3 shows the comparison of typical scales of the SPP energy localization.

An important effect in nanoplasmonics is the resonant excitation of local plasmons (LPs). LP resonance can be observed in metallic particles with typical sizes much smaller than the optical wavelength for which the quasi-static approximation of the electromagnetic response is applicable. The polarizability $p(\omega)$ of a subwavelength particle is proportional to the local field factor [4, 5]:

$$p(\omega) \sim \frac{1}{L_i[\varepsilon_2(\omega) - \varepsilon_1] + \varepsilon_1}, \qquad (2.6)$$

where $\varepsilon_2(\omega)$ is dielectric permittivity of the particle material, ε_1 is a dielectric constant of the environment, and L_i is a factor depending on the shape and size of the particles and the polarization of the incident light ($i = x, y, z$). The denominator of Eq. (2.6) tends to zero at some wavelength leading to a resonance of local plasmon polaritons, i.e., to increase the average kinetic energy of electrons and to an increase in absorption due to scattering of electrons from the lattice of the metal. The coefficients L_i can be calculated as follows:

$$L_x = \int_0^{\infty} \frac{abc\ ds}{2(s+a)^{3/2}(s+b)^{1/2}(s+c)^{1/2}}, \tag{2.7}$$

where a, b, and c are the sizes of the particles in the directions x, y, and z. The equations for L_y and L_z are obtained by cyclic permutation of the variables a, b, and c. The equation $\Sigma L_i = 1$ is fulfilled. The formulae for coefficients L_i for metallic ellipsoid with small cross-section are as follows:

$$L_x = 0, \quad L_y = \frac{c}{b+c}, \quad L_z = \frac{b}{b+c}. \tag{2.8}$$

2.2 Experimental Methods of Excitation of Surface Plasmons

If $\varepsilon_2 < 0$, $\varepsilon_1 > 0$ and $|\varepsilon_2| > \varepsilon_1$, which corresponds, e.g., to the case of a metal in a vacuum environment in the visible, $\mathrm{Re}(k_{\mathrm{spp}}) > \omega/c$. Thus, SPPs cannot be excited by light incident from medium 1. A number of optical schemes can be used to compensate for the difference in wave numbers (see Fig. 2.4).

The schemes with dielectric prisms, such as the Kretschmann or Otto schemes, are commonly used for efficient excitation of plasmon-polaritons (Fig. 2.4a, b). The exciting beam is incident from a medium with a refractive index greater than the index of one surrounding the film. The grating excitation schemes are also widespread (Fig. 2.5). Consider a metal surface modulated with the period of d that is irradiated with light of wavelength λ at an angle of incidence θ. The scattered radiation has a set of diffraction orders. The values of the wave vector projection to the plane of the grating for these orders are $k_x = k_0 \sin\theta + jG = 2\pi\ sin\theta/\lambda + 2\pi j/d$,

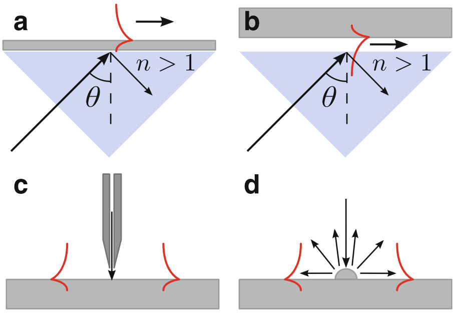


Fig. 2.4 Experimental methods for coupling light to surface plasmons. (**a**) Kretschmann geometry. (**b**) Otto geometry. (**c**) Coupling SPP with a scanning near-field optical microscope tip. (**d**) Surface nanodefect coupling

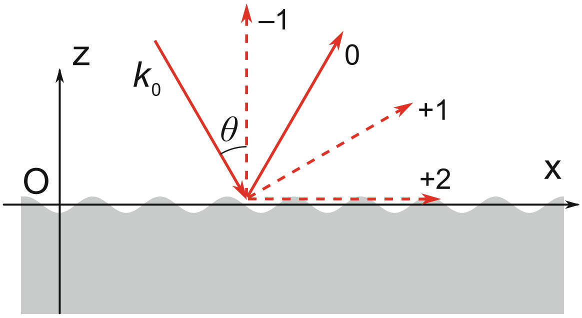


Fig. 2.5 Diffraction of electromagnetic radiation from a metallic diffraction grating and SPP coupling through +2 diffraction order

where G is the grating reciprocal vector, and j is an integer. If the wave number of any diffracted beam coincides with the wave number of SPP:

$$\pm k_{spp} = k_0 \sin\theta + jG, \tag{2.9}$$

the SPP is excited effectively. The central wavelength of the SPP coupled through the grating is written as follows:

$$\lambda = \frac{d}{\pm j}\left(\sqrt{\frac{\varepsilon_2'(\omega)\varepsilon_1}{\varepsilon_2'(\omega)+\varepsilon_1}} \mp \sin\theta\right). \tag{2.10}$$

The condition (2.10) leads to so-called Wood's anomaly [6] which manifests itself as a narrow dip in the reflectance spectra of the metal gratings [7]. Thus, the resonant excitation of surface plasmons in the optical wavelength range in metallic films is possible in the presence of periodic nanostructuring.

2.3 Experimental Methods of Detection of Surface Plasmons

Since SPP is an excitation with the wave vector greater than one in surrounding dielectric, it does not couple to propagating electromagnetic waves. In other words, one cannot use a conventional microscope to observe an SPP by taking an image of the metallic film surface. In order to identify the distribution of plasmonic waves as a function of the coordinate in the plane of the film, it is necessary to use one of methods that are described in this section, namely the leakage radiation microscopy or scanning near-field optical microscopy (SNOM).

The first method of SPP observation relies on the reciprocal character of the Kretschmann excitation scheme [8]. An SPP coupled to the surface of the film via a dielectric prism experiences losses—nonradiative ones and radiative ones. The former stands for ohmic losses due to electron-lattice energy exchange, while the latter emerges from SPPs coupling back to the far-field radiation, e.g., via the same

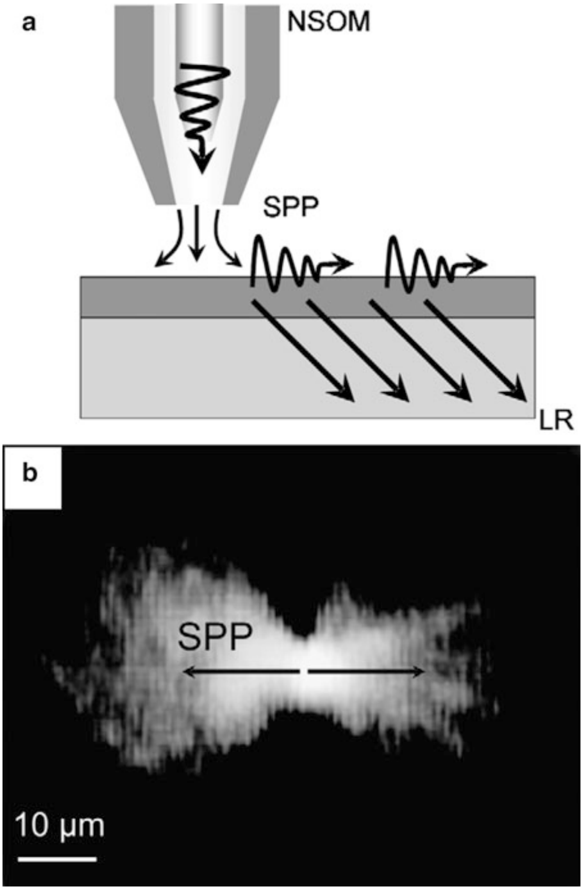


Fig. 2.6 Leakage radiation microscopy image of SPPs launched using a scanning near-field optical microscope tip [8]

channel it was excited. In other words, there would be radiation in the direction of the reflected ray in Fig. 2.4a which is present due to SPPs. Every point of the film subject to carrying SPP energy is giving part of the SPP energy back in proportion to the intensity of the SPP in that point. By taking an image of the film surface though the prism one can directly observe the SPP profile, as shown in Fig. 2.6.

Being one of the conventional far-field microscopy methods, leakage radiation microscopy has significant limitations associated with the diffraction of light. One of the fundamental laws of optics is the existence of the so-called diffraction limit, which determines the minimum distance R between two objects that are said to be resolved from each other using light of wavelength λ:

$$R \simeq 0.61\frac{\lambda}{n},$$

where n is the refraction index of the medium. The limit for the optical wavelength range is of the order of 200–300 nm. Scanning near-field optical microscopy is based on other principles of the image construction, which can overcome the difficulties associated with diffraction of light and realize a spatial resolution of 10 nm and better.

The principles of SNOM lie in the use of a light detector with a size smaller than the wavelength of the radiation. The tip of an optical fiber covered with a thin metal layer with a small hole in it can form such a detector. The detector collects only the portion of the radiation in the vicinity of the hole; so the resolution of this method is determined by the size of the hole which can be made less than 100 nm [9]. The optical image is obtained by scanning the surface of the sample with the probe and constructing a two-dimensional raster of the optical signal value with a resolution exceeding the diffraction limit [10]. Since the distance between the probe and the sample surface is set to be small (starting from 5 nm) and kept constant during scanning, it is also possible to observe and measure the amplitude of the nonradiative excitations of the samples, such as SPP [11], or radiation reflected from the interface between two dielectrics under conditions of total internal reflection [12]. Many properties of plasmonic nanostructures and microsystems have been studied by means of SNOM, including nanofocusing [13], energy transfer along plasmon waveguides by means of SPPs [14, 15], interference of plasmon waves [16, 17], and others.

The same approach of bringing a point-like scatterer to the subwavelength vicinity of SPP is realized by adding subwavelength roughness to the metal surface. Every bump of the film acts like a SNOM tip scattering the SPP energy to the far-field. By taking an optical microscope image of the rough golden film, one can steadily observe the SPP intensity profile. The disadvantage of the method lies in the apparent distortion of the sample initial design as it is usually assumed to be made of a flat metallic film.

2.4 Plasmonic Metamaterials

The development of nanofabrication technology has led to a vast variety of new possibilities of tailoring of optical properties of media. Optical metamaterials are produced from bulk media or thin films by lithographic methods which imply spatial structuring on the scale of tens of nanometers. This gives rise to optical properties absent in the initial material. Negative refractive index [18] and many other noticeable effects were observed recently in optical metamaterials. Noble metals are a common basis for most of the optical metamaterials and these effects are usually associated with the excitation of SPPs. Planar metamaterials based on thin metal films are significantly easier to manufacture than three-dimensional ones and they also manifest a resonant optical response leading to various effects in the optical range. If a periodic array of subwavelength holes is created in a metal film with a thickness of several skin layers, the effect of extraordinary optical transmission (EOT) is observed—the transmitted electromagnetic wave intensity is drastically enhanced at the surface plasmon resonance wavelength in comparison to off-resonance frequencies [19].

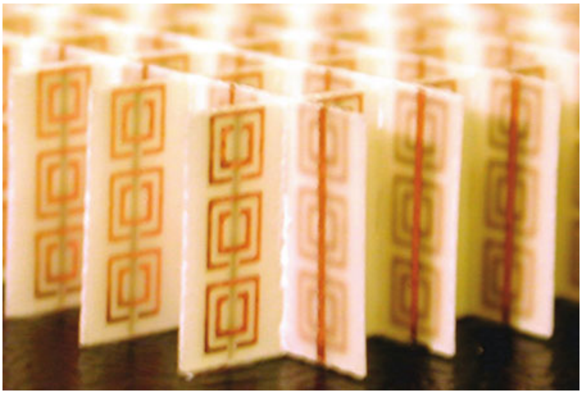


Fig. 2.7 The first experimentally demonstrated negative-index metamaterial. Split current rings induce effective negative permeability and straight vertical rods supply negative permittivity [21]

V.G. Veselago [20] suggested the existence of a hypothetical medium with simultaneously negative dielectric constant ε and magnetic permeability μ for the electromagnetic wave of a given frequency. The Maxwell's equations analysis shows that the wave propagating in a medium with negative *Re*(ε) and *Re*(μ) has the opposite phase and group velocities, which suggests that the refractive index of such media is less than zero. Experimental confirmation of the existence of such media for electromagnetic waves of the radio frequencies [21] required the construction of an array of specially shaped objects (Fig. 2.7) that are resonant for both the electric and magnetic components of the electromagnetic wave. For the realization of media with negative refractive index at optical wavelengths, it is possible to use metals providing a negative real part of the refractive index. However, the bulk magnetization of the media at optical frequencies is close to zero for all known natural materials making the magnetic susceptibility of the latter close to unity. To achieve an effective magnetic response of the medium in the optical range—in other words, to observe the phenomenon of optical magnetism—the nanostructured media can be used, such as ordered arrays of metallic nanoparticles [22]. The electric component of the incident electromagnetic wave can excite both symmetric and antisymmetric oscillations in a pair of nanoparticles. The symmetric oscillations give a nonzero electric dipole moment of the system, while in the case of the antisymmetric resonance the electric dipole moment is close to zero but the counter-flowing "ring" currents yield a magnetic dipole moment emulating the magnetization of the medium. The effective value of the magnetic susceptibility can be different from unity, as well as negative value, which opens the way for the practical realization of materials with negative refractive index [18, 23].

Another anomaly was observed in optical metamaterials made of thin films perforated with an ordered array of subwavelength apertures. According to Bethe's model [24] of light transmission through the circular subwavelength aperture of radius *r* made in the thin conducting film, the transmission coefficient can be written as follows:

$$T(\lambda) = 64(kr)^4/27\pi^2, \tag{2.11}$$

where $k = 2\pi/\lambda$ is the wavenumber. Later on, it was shown [19] that the regular array of subwavelength apertures fabricated in non-transparent silver film and arranged with the minimal period of d possesses the transmission coefficient spectrum having peaks at the wavelengths corresponding to SPP resonances described by Eq. (2.10). The maximal value of transmission was shown to be approximately two orders of magnitude larger than that expected by Eq. (2.11). This phenomenon is now known as the extraordinary optical transmission effect since the total light transmission normalized at the aperture area appears to be larger than unity. One of the mechanisms responsible for electromagnetic energy transfer from one side of the nanoperforated metallic film to another one is the resonant SPP excitation at the film surface followed by penetration of SPP modes through the nanoholes and further coherent diffraction of SPP at the film backside into the propagating electromagnetic wave according to the phase matching conditions at the periodic array of nanoholes [25, 26].

One of the spectacular features of the EOT effect is the maintenance of spatial coherence of the collimated beam upon transmission through the nanohole array supported by recent experiments on entanglement of a biphoton as one of its parts was transmitted through the nanohole array [27]. Small losses of the entanglement show the quantum nature of the EOT phenomenon that allows consideration of SPP as quasiparticles with a lifetime inversely proportional to the spectral width of the SPP resonance.

The EOT effect possesses strong spectral selectivity for the incoming radiation since the spectral position of the EOT peak corresponds to the spectral line of the SPP excitation at the surface of the nanoperforated sample [3]. The spectral position of the EOT peak can be tuned by changing the geometrical parameters of the structure such as the nanoholes period, form and size of holes, and film thickness. Therefore, the optical properties of periodically modulated surfaces of noble metals are also defined by the effectiveness of the excitation of the SPP modes. However, since electromagnetic radiation induced by coherent SPP scattering at the spatially modulated surface can interfere with the incoming light exciting surface plasmons, the spectral lineshape of the response of such structures can significantly differ from the Lorentz lineshape.

The discovery of EOT [19] in 1998 yielded long and intensive debates of its nature and different interpretations of the effect. The main problem was that the calculated surface plasmon resonances did not coincide with experimental EOT peaks. The anomalous increase in transmission both of subwavelength hole arrays and gratings with narrow slits can be caused by the excitation of SPPs at the two interfaces of a film, as well as of the waveguide plasmon modes within the holes [28]. This interpretation was confirmed in further experiments: it was shown that the resonant transmission through a one-dimensional lattice of 60-nm-thick gold film can be caused by SPP excited at the upper interface only, as well as by localized waveguide modes [29]. On the other hand, Cao and colleagues have obtained a decrease in transmittance at the of SPP excitation wavelength for thicker films of 4 μm, and argued that the EOT arises only due to the localized waveguide modes [30]. Later, it was assumed that the localized plasmons at the holes play an important role in EOT. A lot of different shapes of holes were considered and it was found that their shape and configuration are important for the fixed period [31]. The more complex is the shape of a hole, the easier nonpolarized light couples into its local mode. Moreover, the symmetry of two-dimensional lattices is important for transmittance value. For example, it was found that EOT was stronger for hexagonal lattice than that for square lattice [32]. Then it became clear that the excitation of SPP affects the resonance characteristics of localized plasmons. For example, if Wood's anomaly appears, the resonances of localized plasmons are red-shifted due to Fano-type interference. It was shown for very narrow slits (much less than the thickness of the film) that the resonance of SPP reduce the transmission. The situation can be completely different for slits width comparable to the period, because local modes inside the slits cannot be excited any more, while local plasma oscillations can appear within the stripes.

2.5 Optically Anisotropic Metamaterials: Versatile Polarization Control with Plasmonic Nanostructures

2.5.1 Far-Field Polarization Control

For changing the state of polarization (SoP) of light two key optical elements are commonly used, namely a polarizer and a wave plate. While polarizers are used to prepare the output SoP which does not depend on the input SoP, wave plates are used to transform the incident SoP. The principles of operation of a polarizer which prepares a linearly polarized state rely on the effect of linear dichroism under which only linearly polarized light is transmitted through and the orthogonal polarization is reflected, deflected, or absorbed. The wave plate introduces a phase delay between two orthogonal linear polarizations which are eigenpolarizations of the wave plate; as a result general elliptical polarization is the output from the device. Since there are two parameters defining any fully polarized SoP, i.e., the ellipticity E of the polarization ellipse and its major axis orientation with respect to the laboratory frame, φ, there are also two parameters which define the SoP output from any wave plate. These parameters are wave plate optical axis orientation angle ψ and the amount of phase delay introduced between the eigenpolarizations, $\Delta\phi$.

In comparison to the bulk polarization elements plasmonic nanostructures comprise materials of essentially subwavelength thicknesses. Thus the thickness cannot be continuously tuned to achieve the desired phase shift between the eigenpolarizations. Nevertheless plasmonic anisotropic nanostructures were proven to compete with the bulk polarization optics [33–37]. In this section we investigate what parameters could be tuned in order to attain the desired polarization state output from an anisotropic plasmonic nanostructure.

The sensitivity of SPPs to the polarization of light allows for the observation of linear birefringence and dichroism. Consider a metal surface that is modulated along the Ox direction and radiation impinging on the surface at an angle θ in the plane of incidence parallel to Ox (Fig. 2.5). It is possible to excite an SPP according to phase-matching conditions (2.10) with the in-plane polarized light. A dip will be observable in reflection spectra corresponding to energy transfer from the incident wave to SPP. Light polarized perpendicularly to the plane of incidence (s-polarization) is not capable of SPP excitation and is reflected without significant losses. Due to symmetry restraints s- and p-polarized waves are eigenpolarizations of the grating since the polarization state is not changed by reflection. An SoP, which comprises a linear combination of these eigenstates, may undergo changes depending on so-called ellipsometric constants of the medium:

$$\xi = \left|\frac{r_p}{r_s}\right|, \quad \Delta\varphi = \varphi_p - \varphi_s, \tag{2.12}$$

where $r_{5p} = |r_p| \exp i\varphi_p$ and $r_s = |r_s| \exp i\varphi_s$ are field reflection coefficients of p- and s-polarized waves, respectively. Since the effectiveness of SPP excitation has strong dispersion, a pronounced spectral inhomogeneity is observed for ξ and $\Delta\varphi$.

Dispersion and, hence, polarization transformation performance depend on the parameters of the resonance. The spectral response of the most resonant physical systems is described by a Lorentzian line shape. The spectrum of a system response to harmonic excitation of frequency ω is described by a classical harmonic oscillator model:

$$T_L(\omega) = \frac{B}{\omega^2 - \omega_0^2 - 2i\gamma\omega}, \tag{2.13}$$

or with the assumption of $\omega_0 \gg \gamma$ for analytical signals :

$$T_L(\omega) = \frac{B}{\omega - \omega_0 - i\gamma}, \tag{2.14}$$

where B is the amplitude of the excitation, ω_0 the natural frequency of the oscillator, and γ the damping factor. However, the observed signal is often a mix of T_L and the exciting signal. In this case the spectral line becomes asymmetric of the Fano-type [7]:

$$T_F(\omega) = A + \frac{Be^{i\phi}}{\omega - \omega_0 - i\gamma}, \tag{2.15}$$

where A is the amplitude of nonresonant signal, and ϕ the phase between resonant and nonresonant signals. The shape of the Fano resonance is a common lineshape of the optical response spectra of nanostructured systems supporting the excitation and propagation of SPP. Thus, the transmittance spectra of thin films of metal with a periodically modulated surface contain Fano resonance features. This is the consequence of the fact that the reflected signal contains two coherent components: the light transmitted through the film directly, and the radiation coupled into a surface mode and then diffracted in accordance with the phase-matching condition (2.10). Fano resonances in the optical response of nanostructures are attracting considerable interest due to their high spectral selectivity [3, 38].

By using a plasmonic nanostructure with a Fano resonance, one can attain high polarization conversion by means of plasmon-enhanced birefringence and dichroism [36]. One may use nanoslit arrays in a thin gold film to achieve a sharp Fano-type resonance for visible wavelengths. The transmission of such a sample is presented in Fig. 2.8a as a function of incident light's photon energy and wave vector projection onto a sample's plane k_x for the *p*-polarized light. A sharp and angle-dependent Fano-type resonance is observed in the visible range; a cross-section of the transmission function at the angle of incidence $\theta = 50°$ is shown in Fig. 2.8b. The origin of the Fano-type optical response lies in the coherent superposition of the grating SPPs resonance response and the background signal directly transmitted through the 30-nm-thick gold film. The phase delay $\Delta\varphi$ between E_y and E_x of the output *E*-field's components and the dichroism $|E_x|^2/|E_y|^2$ are measured in the spectral domain and shown in Fig. 2.9a for $\theta = 50°$. The phase difference varies from 0.38π to 0.85π, which could be obtained by a medium of the same thickness with extreme ordinary–extraordinary refractive indices differences of $\Delta n \simeq 4.4 - 10.4$.

Note that the spectral position of a feature in the $\Delta\varphi$ spectrum is connected to the position of the SPP resonance. It means that if one fixes the operating wavelength, there is a strong dependence of the phase delay on the angle of incidence. The phase delay could be tuned continuously within the aforementioned range of 0.38π–0.85π by varying the angle of incidence within the range of 46° to 54°. The angle of incidence stands as a birefringence control parameter—a substitute to the wave plate thickness in the bulk polarization optics.

2.5.2 Near-Field Polarization Control

Surface plasmon-polaritons being related to spatial electron plasma oscillations in a metal are strongly localized to nearby the metal surface. The characteristic scale of the localization of bounded electromagnetic waves depends on various factors such as the form of metallic objects and curvature of their elements and appears to be from several nanometers to subwavelength scale, which in a majority of cases is sufficiently lower than diffraction limit for the optical wavelength used. Concentration of electromagnetic energy in the objects with the a size smaller than the optical wavelength in vacuum, which is called subwavelength focusing or superfocusing, and optical signal transfer over subwavelength channels were experimentally demonstrated in various systems, for example, in individual nanoparticles [39], in nanoparticle chains [40, 41], in V-shaped subwavelength grooves fabricated at the smooth metallic surfaces [13], and at near-field optical microscopy probes [42], involving interference of standing plasmonic waves [43], utilizing radially [44] or

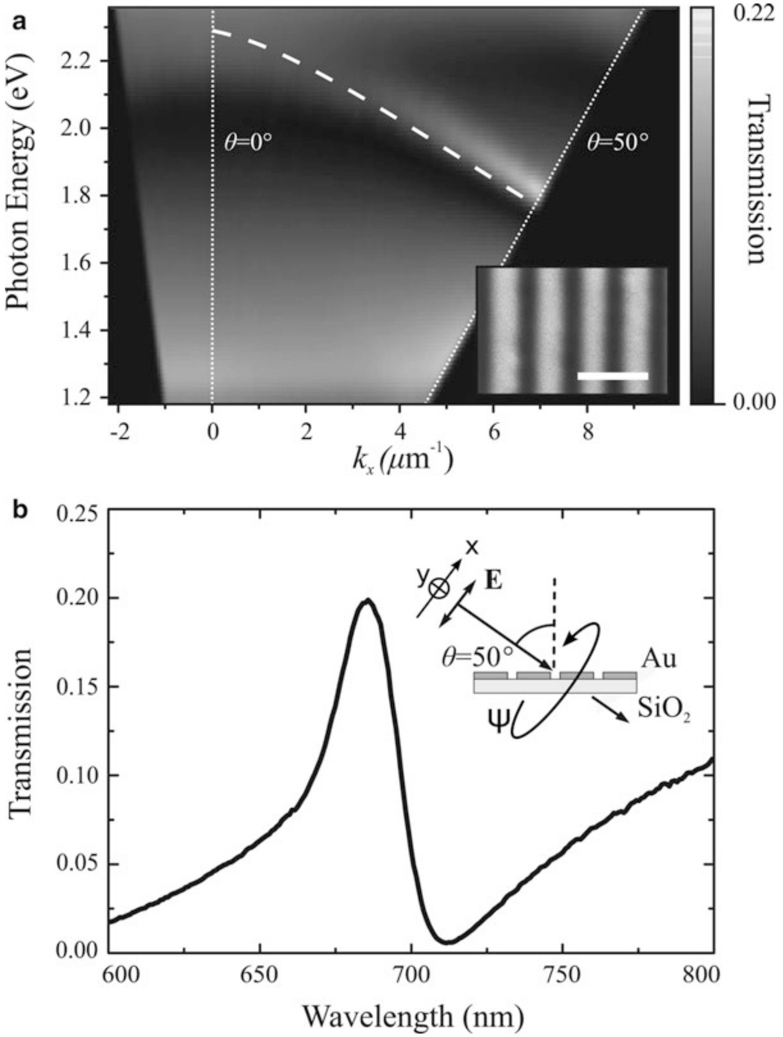


Fig. 2.8 (**a**) Transmission coefficient of the nanoslit sample as a function of photon energy and transversal wavenumber of the incident *p*-polarized light. SPP excited with the blazing − 1 diffraction order is present. The dispersion relation of SPP propagating at a gold-fused silica interface estimated using the expression (2.10) is denoted with the *white dashed line*. The inset shows the SEM picture of the sample, the bar length equals 500 nm. (**b**) The transmission spectrum at the 50°-incidence indicating a Fano-type resonance for the *p*-polarized light [37]

circularly [45] polarized beams. The practical realization of concentrating electromagnetic energy at the scale smaller than the optical wavelength in vacuum seems to be an important task towards miniaturization of optical devices such as nanolasers [46, 47] or elements for optical computing [48].

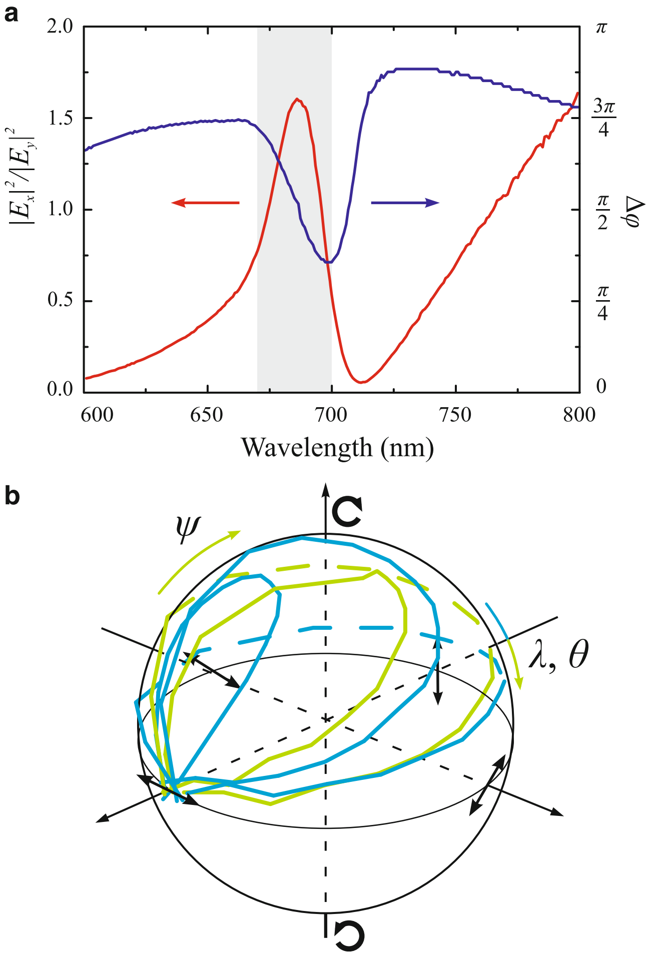


Fig. 2.9 (**a**) Linear birefringence and dichroism of the nanoslit sample represented by the phase retardation $\Delta\varphi$ and $|E_x|\ ^2/|E_y|^2$ spectra, respectively, measured at the 50°-incidence. (**b**) The Poincaré sphere representing the map of experimental polarization transformations done by the system under study. The input state is horizontally polarized. Each curve describes a set of output states for azimuthal angle Ψ varying from 0° to 90°. Different curves stand for different wavelengths in the vicinity of Fano resonance or different angles of incidence, calculated using the $\partial\theta/\partial\lambda$ expression from the text. Although not indicated, the bottom hemisphere is covered in the same way with $-\ \Psi$ angles due to symmetry relations [37]

Using the fact of subwavelength electromagnetic energy localization with SPPs, we show that not only energy density but also polarization could be separated in space by plasmonic nanostructures. To do this we measure linear dichroism in the subwavelength vicinity of an anisotropic plasmonic nanostructures. The nanostructures under study comprise very long plasmonic nanoparticles with a local plasmon resonance with the central wavelength of 580 nm excited with the

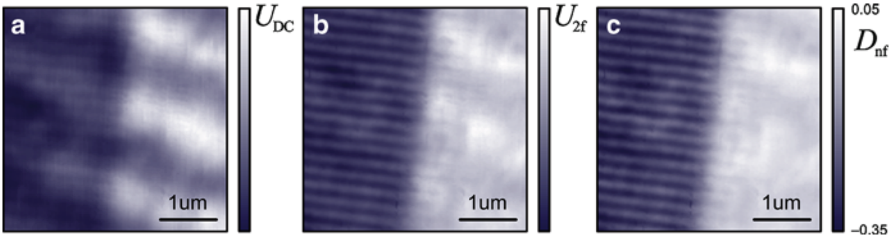


Fig. 2.10 (**a**) Distribution of U_{DC} in the subwavelength vicinity of the sample. (**b**) Distribution of U_{2f} in the subwavelength vicinity of the sample. (**c**) The ratio of the latter to the former showing the linear dichroism map

linear polarization perpendicular to the stripes. Near-field polarization properties of plasmonic nanogratings were investigated by means of scanning near-field dynamic polarimetry setup [49]. Horizontally polarized light from a doubled CW Nd:YAG laser with $\lambda = 532$ nm passes through a half-wave plate with its axis oriented at 22.5° with respect to the optical table. Diagonally polarized light passes through a photoelastic modulator (PEM) that temporally modulates the phase shift between vertically and horizontally polarized components. The phase shift undergoes harmonic oscillations $\phi(t) = A\sin(2\pi ft)$, where A is the phase shift oscillations amplitude and f is the PEM operating frequency which is equal to 47 kHz. If $A = \pi$ the polarization state output from the PEM changes from diagonal to antidiagonal with the $2f$ frequency. A half-wave plate placed behind the PEM is used for transformation of the polarization modulation from diagonal/antidiagonal to horizontal/vertical. Mirrors are used to deliver the light onto the sample through the substrate. Optical signal is then collected by an aperture SNOM fiber probe with the aperture diameter of 50–100 nm. The distance between the probe and the sample is about $\lambda/20$ and controlled by a three-coordinate piezo-driver. The collected signal is sent to a photomultiplier tube (PMT) and then divided into two channels. The first channel is connected to a lock-in amplifier which detects the signal modulation amplitude U_{2f} at the $2f$ frequency which is caused by the linear dichroism. The second channel is a low pass filter with the cutoff frequency of 600 Hz. Consequently both signals, U_{2f}, which is a measure of the linear dichroism, and U_{DC} which is proportional to transmittance, are simultaneously measured.

An example of the maps of U_{2f} and U_{DC} signals are given in Fig. 2.10a,b. Now obtaining the map of the absolute value of linear dichroism consists of dividing the former by latter [50, 51]. As a result we obtain a clear, speckle-free map of linear dichroism in Fig 2.10c which shows one how the polarization is distributed in the vicinity of the nanostructure. In other words, there are different places in the vicinity of the sample which are separated by a distance of 150 nm where light is polarized differently if the incident light state is not horizontally or vertically polarized one. While the mean dichroism value in the near-field regime -0.21 ± 0.03 coincides with one in the far field -0.20 ± 0.02, there is a subwavelength distribution of it in the plane of the sample.

2.6 Components of Nanoplasmonics for On-Chip Integration

In this section we examine the possibilities of creating basic components of the so-called plasmonic circuitry, which allow for plasmonic signal manipulation. The plasmonic circuit elements are divided into passive ones, which control the distribution of plasmonic energy corresponding to their shape, and active ones, which introduce the possibility of modulation of plasmonic signal by means of external electric or magnetic fields or electromagnetic radiation.

2.6.1 Passive Elements

The injection of surface plasmons onto the surface of a metallic film could be done by using methods described in Sect. 2.2. However, it is desirable to reduce the size of the photon-to-plasmon conversion terminal; for this, a single subwavelength defect in the surface of the film could be used. Figure 2.11 illustrates the distribution of the electromagnetic field density in the vicinity of the nanohole milled in a gold film illuminated with a monochromatic linearly polarized plane wave. Surface waves are seen spreading from the center of the aperture forming a dipole-like cosine-squared angular radiation pattern [52].

Figure 2.12 illustrates how an array of such apertures could be used as plasmonic field concentrators [53]. Each aperture is a point-like source of SPPs, which interfere in a preferential direction along the axis of mirror symmetry of the structure.

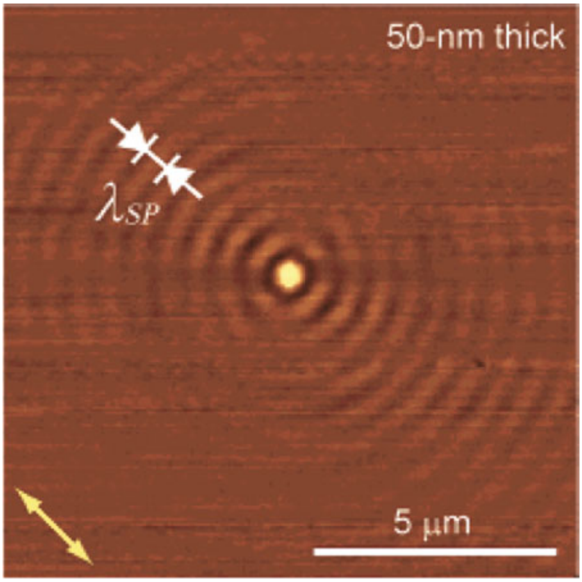


Fig. 2.11 The source of SPPs: electromagnetic field density in the vicinity of a 50-nm-thick golden film as measured with a SNOM. SPP Radiation pattern is seen with the main direction along the polarization of the incidence denoted with an *arrow* in the *bottom-left* part of the panel [52]

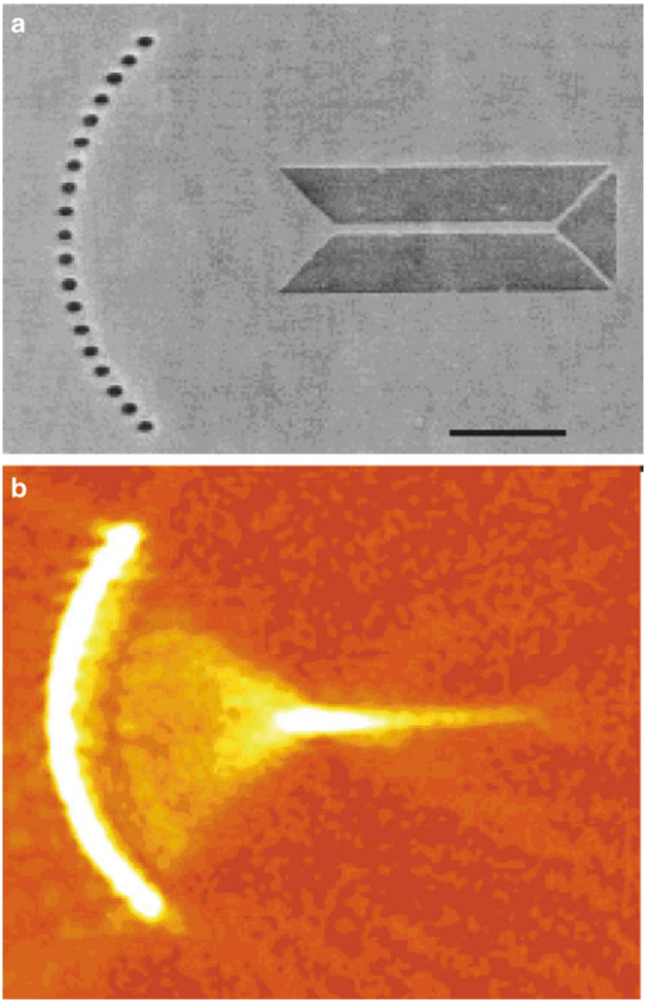


Fig. 2.12 Coupling of SPPs to waveguides. (**a**) An SEM image of an array of subwavelength apertures (*left*) and a ridge waveguide (*right*). (**b**) Focusing of plasmonic signal emitted from the apertures into the waveguide [53]

The result is a plasmon coupled to the bifurcation of the ridge waveguide with the width of 100 nm which scales as λ/5 at the wavelength used in the experiment.

The plasmonic hot spot in Fig. 2.12 is located at the beginning of a plasmonic waveguide that comprises a strip of gold along which SPPs can propagate. These so-called ridge waveguides allow for propagation of SPPs since SPP is a wave confined to the metal surface and follows the shape of it. However, SPPs undergo dramatic radiation losses in these waveguides since SPP is scattered by its edges. There are several possibilities to suppress these losses.

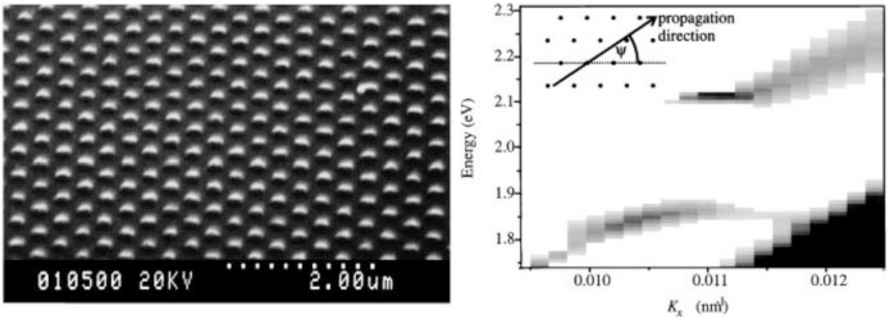


Fig. 2.13 *Left*: scanning electron micrograph of a plasmonic crystal surface. *Right*: visualization of the plasmonic band gap by measurement of the frequency and angular domain reflection spectra in the Kretschmann experimental geometry [54]

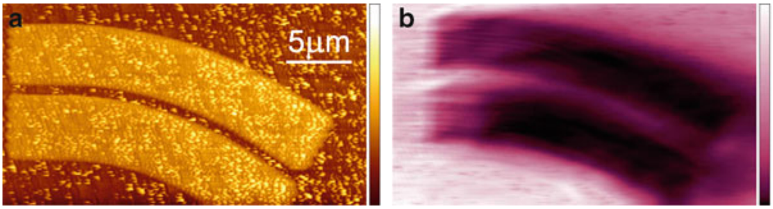


Fig. 2.14 SPPs in plasmonic crystal waveguides. (**a**) Topography of a plasmonic crystal-based bent waveguide. (**b**) SPP intensity distribution in the vicinity of the waveguide as measured by SNOM [55]

One of them is the usage of bandgap (BG) structures. A BG plasmonic structure is a periodically modulated metal-dielectric interface illustrated by the example of dielectric pillars on top of the metallic film in Fig. 2.13 [54]. When a plasmonic wave encounters such an array of scatterers, the secondary plasmonic waves may interfere in a way that no signal is transmitted through the BG structure. In this case a straight array of missed pillars acts as a waveguide for SPPs since the field in the areas of pillars tends to zero. The BG waveguides were demonstrated to guide SPPs along bent channels with low bending losses, as shown in Fig. 2.14.

The best plasmonic waveguide design in terms of low radiation losses is the so-called channel waveguides that represent a deep subwavelength groove in a metal film [13]. For the basic mode of the waveguide the SPP field is concentrated on the very end of the ditch making it hard to couple to the far-field and thus giving minor radiation losses as SPP is propagating along it. In Fig. 2.15c,d an experimental realization of such a waveguide is performed in a metal film. A 90° low-loss bend could be achieved by proper impedance matching of the waveguides it incorporates (Fig. 2.15a,b).

One can split one plasmonic wave into two or combine two of them into one by using a splitted waveguide as shown in Fig. 2.16. The topography of the surface of

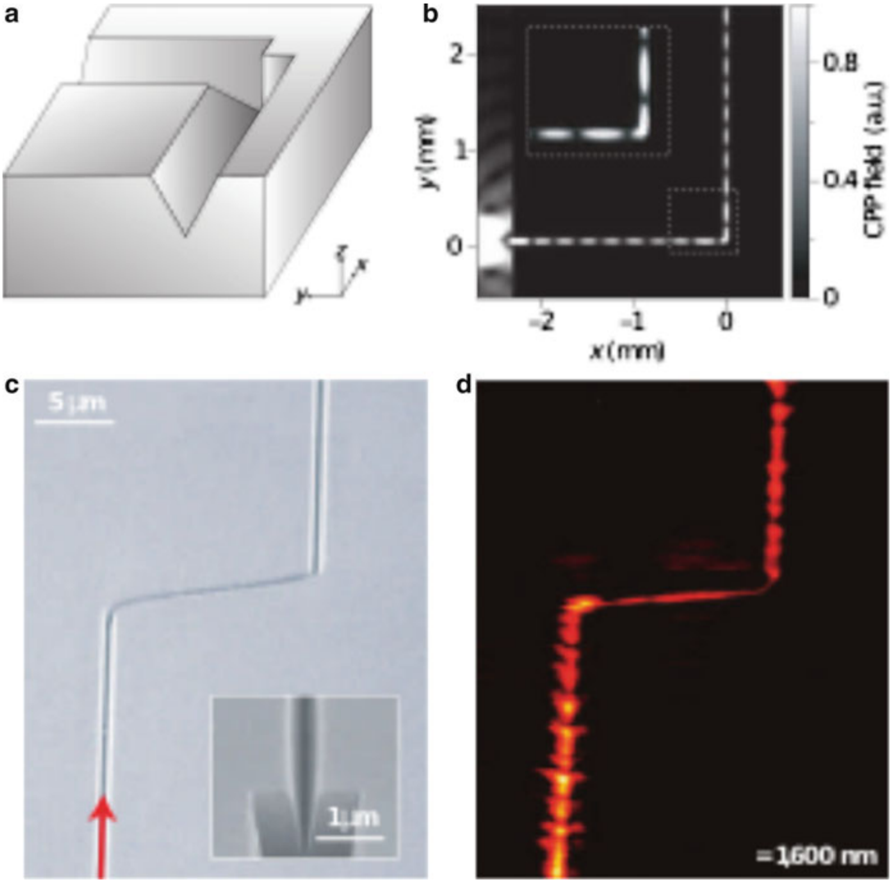


Fig. 2.15 SPPs in the V-groove waveguides. (**a**) Design of a 90°-bent channel plasmonic waveguide. (**b**) Results of calculation of the electromagnetic field intensity in the waveguide. (**c**) A scanning electron micrograph of a channel waveguide with a twist. (**d**) Respective SNOM image of SPP traveling along the waveguide [13]

the device is shown along with a near-field optical microscope image demonstrating field localization within the waveguide and its division over two plasmonic channels after the splitter.

Two plasmonic splitters could be combined into a Mach–Zehnder (MZ) interferometer. It has one input and one output; the intensity of the output signal depends on the phase of the waves which interfere within it. If one of the waves is retarded with respect to another by means of externally changing the local refractive index of gold or surrounding medium in one of the MZ arms, the output intensity would differ from one without this change. Thus the plasmonic output from such a device could be controlled via external stimuli allowing for modulation of the plasmonic signal for communication. So-called active ways of plasmonic signal creation and modulation are considered below.

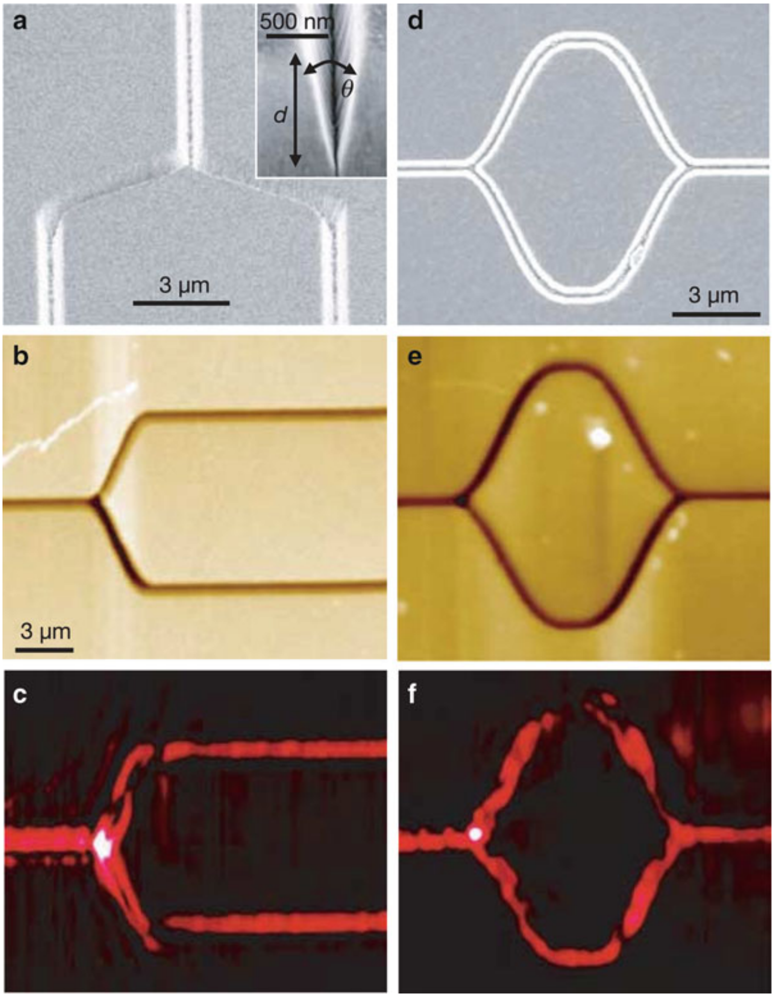


Fig. 2.16 (**a**)–(**c**) A channel-waveguide-based plasmonic splitter as viewed in a scanning electron microscope, atomic-force microscope, and SNOM. (**d**)–(**f**) A channel-waveguide-based plasmonic Mach–Zehnder interferometer as viewed in a scanning electron microscope, atomic-force microscope, and SNOM [13]

2.6.2 Active Elements

Like commercially available fiber-optic networks, every optical circuit requires a source of coherent radiation to be processed. Lasers are commonly used in optic networks; yet, the source of a coherent surface plasmon signal—a surface plasmon laser or a spaser—has been a long-standing problem for the community [47, 56–61]. The general idea of a spaser is depicted in Fig. 2.17. The active medium of a spaser is a luminescent substance—usually a dye or semiconductor quantum dots—that is placed in the near-field vicinity of a plasmonic nanoparticle, which

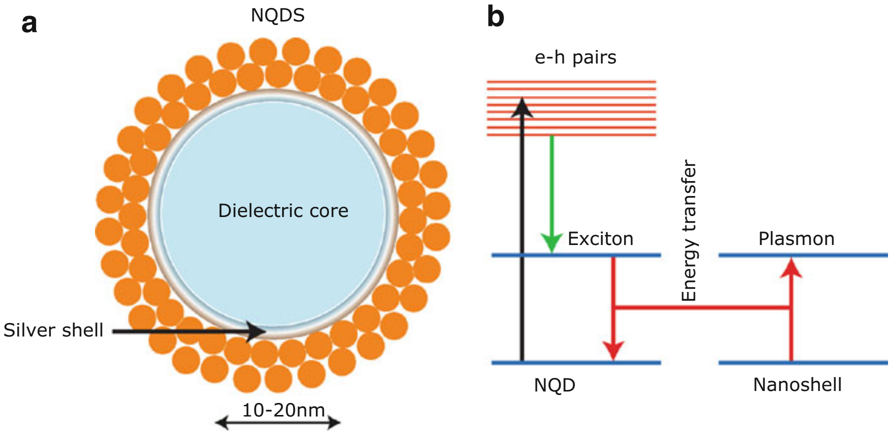


Fig. 2.17 A concept of a spaser. (**a**) A dielectriccore-shell dielectric-metal nanoparticle covered with nano quantum dots (NQDs). (**b**) Energy level diagram of the spaser. The *leftmost arrow* demonstrates the pumping of the NQDs with external radiation [56]

acts analogously to a cavity of a conventional laser. Luminescence is induced by optically pumping the active medium. After the pump power reaches certain critical value, the system starts to emit coherent surface plasmons within the spectral line of the nanoparticle's local plasmon resonance in the stimulated regime. Such a source could be placed inside a plasmonic waveguide since its dimensions are much smaller than a wavelength of an SPP.

An experimental verification of the possibility of a spaser was reported recently [57]. It was shown that a suspension of golden nanoparticles covered with a dye-doped silica shell demonstrates stimulated emission of photons in the far-field, which is a fingerprint of stimulated emission of SPPs in nanoparticles.

In order for information to be transferred via an SPP channel the latter should be modulated. Several methods of amplitude modulation of SPPs have been proposed. Applying an electric field could be used to modulate the SPP amplitude in a device called a metal-oxide-semiconductor field effect plasmonic modulator [48]. It consists of a multilayered structure with an optical source, a drain, and an electric gate (see Fig. 2.18). A surface plasmon mode is specific to an interface between metal and SiO_2 layers; an optical mode is specific to a Si waveguide; applying a gate voltage one sets the quality of the optical mode that may interfere with SPP giving rise to modulated transmission of the device as the gate voltage is modulated.

Another way is to control SPP by means of an optical signal, which is feasible for THz-clock systems. An example of ultrafast control over the SPP coupling to a gold film is demonstrated in a pump-probe experiment on plasmonic gratings [62]. Two sources of ultrashort laser pulses are used. The beam from the first one called the probe is coupled to the SPP mode by the grating coupling method described in Sect. 2.2. The other source provides a more powerful beam called the pump, which illuminates the sample and optically changes the dielectric properties of the dielectric surrounding the grating. Since the SPP resonance position strongly depends on

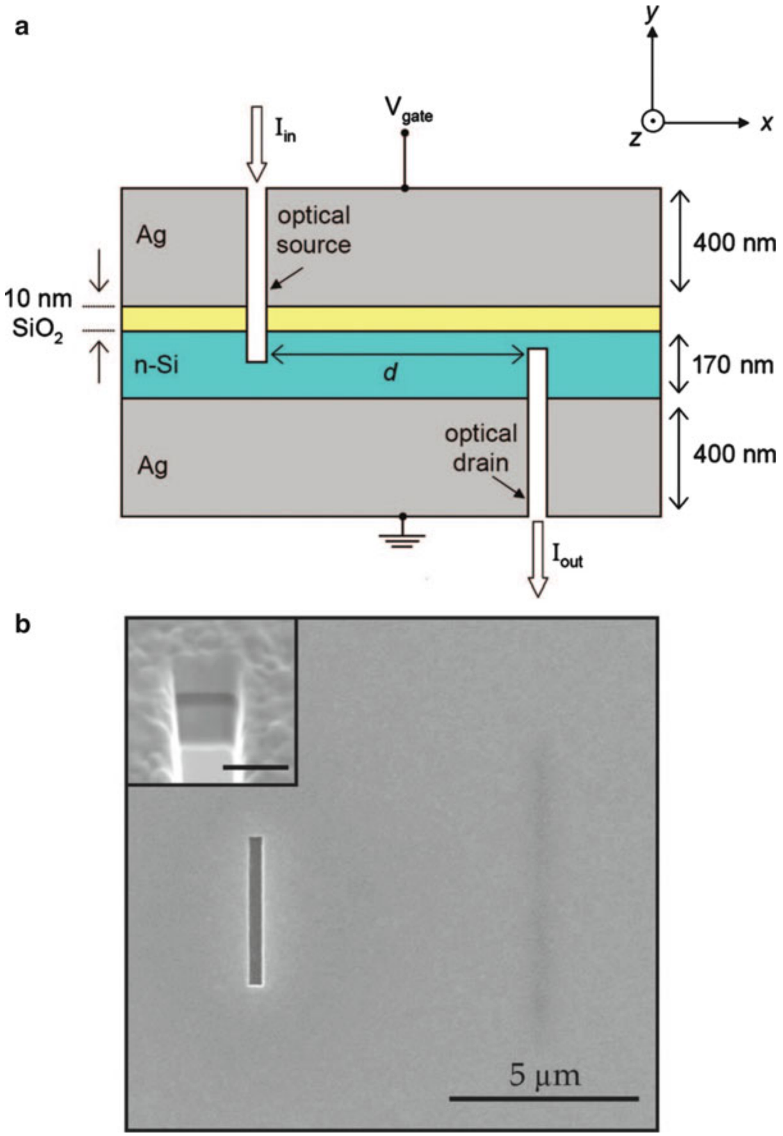


Fig. 2.18 (**a**) A schematic of a metal-oxide-semiconductor field effect plasmonic modulator. (**b**) A scanning electron microscope image of a prototype [48]

the dielectric properties of the surrounding medium [see Eq. (2.10)], even small changes of the dielectric permittivity of silicon induced by the carrier emission cause a spectral shift of the SPP resonance. Consequently, the SPP amplitude is diminished when the pump is present if the probe is at the SPP resonant wavelength in the unperturbed case. The speed of such switching lies at the sub-picosecond timescale, which paves the way for terahertz communications.

2.7 Ultrafast Nanoplasmonics

Nonequilibrium electron dynamics in metal nanoparticles and at plain and periodically perforated metal surfaces have been intensively studied during last decade using various femtosecond pump-probe and correlation spectroscopy techniques. The resonant enhancement of the transient thermoreflectance at unstructured surfaces due to surface plasmon excitation is shown to be a sensitive tool for time-resolved measurements of the metal surface temperature at the nanosecond [63] and subpicosecond [64] timescales. For the plane surfaces the SPP lifetime depends strongly on the metal optical constants and the surface roughness. The basic motivation of the studies of the SPP dynamics in nanostructures is the significant changes in the SPP relaxation times due to radiative damping suppression caused by the resonant excitation of the surface or localized plasmon modes. The ability to control these changes and to maximize the plasmon relaxation time for optical applications is an important motivation for this set of problems. For example, a significant reduction of both radiative and nonradiative plasmon damping was observed in gold nanorods in comparison to that in gold nanospheres of the similar sizes [65]. The resonance widths for individual particles of different sizes and shapes were measured. In the nanorods the damping is much slower and Q-factor is several times greater than that in the nanospheres. The observed difference is the strongest for the particles with red resonances, i.e., spheres with a larger size (high damping) and for nanorods with a larger length to diameter ratio (low damping). The local-plasmon dephasing times range from 1.4 to 5 fs for nanospheres and from 6 to 18 fs for nanorods as deduced from the experiment. The maximum dephasing time of 18 fs achieved for nanorods cannot be caused by changes in the radiative decay rate since its contribution is small for the small volume of particles. The reduced nonradiative decay in nanorods is explained by the fact that a threshold energy of about 1.8 eV in gold is higher than the resonant energy of the excited plasmons. For gold nanospheres this effect is screened by the strong radiative damping for the particular spectral region. Analysis of the results shows that contributions of radiative damping, interface damping, and pure dephasing to the process of dephasing are small in gold nanorods. The term “pure dephasing” means some additional elastic phase-loss process of the collective excitation.

The most appropriate technique for studying SPP dynamics is the optical femtosecond cross-correlation technique, which has many modifications. The common principle is the measurement of the correlation function (CF) of the reference femtosecond Gauss-shaped pulse and the pulse modified by the reflection from the studied object. The CF gives the time-resolved information of the processes occurring in the sample. The SPP relaxation processes described above depend strongly on the wavelength, thus a spectral modification of the cross-correlation techniques is required. A series of phase-sensitive correlation techniques is useful, such as frequency-resolved optical gating (FROG) and spectral interferometry. The methods give full parameters of the modified femtosecond pulse including the amplitude and phase of the electromagnetic field at any moment of time.

The metallic surface structuring allows the control of the surface plasmon lifetime for ultrafast plasmonics applications because SPP lifetime in plasmonic crystals can significantly vary depending on excitation and backscattering conditions (i.e., phase relations) as well as effective dispersion modification and photonic band gap presence within the spectral region of interest [66]. The strong distortion of a femtosecond pulse shape is reported in [67] due to resonant surface plasmon excitation in one-dimensional planar plasmonic crystal. The lifetimes of the SPPs excited on the lower and higher band gap edges differ as 18-fs at low-frequency band-gap edge and 250-fs at high-frequency band-gap. The competition of two relaxation channels gives rise to significant variations of the lifetime. The first channel is electron–electron and electron–phonon relaxation with the typical timescale similar to that at smooth surfaces. The second, which is much faster, is the radiative channel. Resulting lifetime value depends on the two channels' efficiencies ratio. The efficiency of the SPP reradiation in specular direction depends on the effective dispersion law slope and the SPP excitation method. The diffraction grating excitation method gives rise to the Wood's anomaly in the reflectance spectrum in the form of a Fano-type resonance. In this case, the radiation relaxation efficiency is strongly spectrally dependent on the vicinity of the resonance due to interference effects.

Pronounced light polarization conversion can be observed in thin metallic gratings due to strong optical anisotropy [37]. Moreover, the polarization conversion occurs at subpicosecond timescale, which was experimentally shown in [68] by using an original polarization-sensitive femtosecond correlation technique. One-dimensional plasmonic crystals were proposed as promising compact media for ultrafast polarization control as an application of the observed effect.

The ultrafast dephasing time of waveguide-plasmon polariton, which is a quasi-particle emerging due to strong coupling between the waveguide modes and the local plasmons, was studied in a 2D metallic photonic crystal structure with a nonlinear autocorrelation technique [69]. A phase-sensitive experimental setup with a stabilized Michelson interferometer and a 13-fs Ti:Sapp laser were used. A prolonged dephasing time of waveguide-plasmon polaritons is shown in metallic photonic crystals when compared to an undisturbed local plasmon. The experimental results demonstrate that it is possible to tailor the dephasing of the coherent excitation in metallic photonic crystals by tuning the coupling strength between the electronic and photonic resonances.

Coherent coupling between optical excitations can lead to a significant pulse propagation delay and a strong modification of the radiative damping rate [70–72]. By using an original time-resolved interferometric technique, a 6-fs delay of a 100-fs pulse transmitted through subwavelength hole arrays was observed [70]. The effective group velocity of $c/7$ of the pulse propagation through the media is measured. The experimental evidence of the subradiant damping in one-dimensional plasmonic crystal is given in [71]. The SPP lifetimes longer than 200 fs are demonstrated. Tuning of the SPP dephasing is demonstrated [73, 74]. In two-dimensional Au hole arrays, the reflectance linewidths decrease and the SPP lifetime increases with decreasing hole size and increasing wavelength [75].

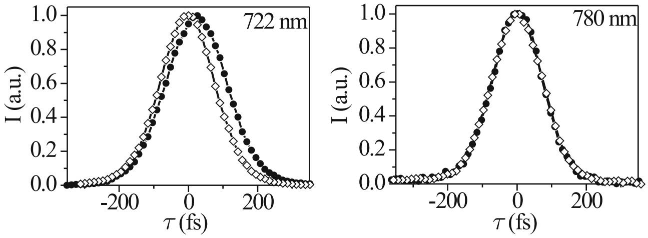


Fig. 2.19 Normalized second order intensity correlation function for p-polarization (*solid circles*) and s-polarization (*open circles*) at $\lambda = 722$ nm and $\lambda = 780$ nm [76]

A temporal modification of femtosecond pulses upon the resonant excitation of the SPPs was studied in one-dimensional metallic nanogratings by femtosecond correlation spectroscopy for the laser pulse duration comparable to the SPP lifetime [76]. The modification reveals itself as the time shift and pulse duration changes. Spectral dependence of the pulse distortion is described by the Fano resonance line shape. The grating was a 50-nm-thick silver film deposited onto a polymer substrate with a one-dimensional periodic surface topography modulation. The cross-section of the sample was close to a sinusoidal profile with the period of 750 nm and the modulation depth of 60 nm. Three resonances in the frequency-angular spectra were observed at three diffraction orders with $j = 1, \pm 2$, and ± 3 according to Eq. (2.10). The plasmonic origin of the resonance features observed in the vicinity of 500 nm and 725 nm was confirmed by the absence of such features in the s-polarized light reflectance. The interference of the reflected light and reradiated SPP yields Fano-shaped resonances (see Sect. 2.5). The spectral line shape of the reflectance is a sum of the nonresonant reflection of incident radiation and the resonance profile of SPP with the Lorentzian line shape as stated by Eq. (2.15). The SPP decay time calculated from the resonance width Γ is $t_{\text{spp}} \simeq 90$ fs. The temporal modification of the femtosecond pulses was studied by the correlation spectroscopy. A Ti: sapphire laser with a pulse duration of approximately 200 fs, a repetition rate of 80 MHz, an average power of 100 mW, and the output wavelength tunable from 690 to 1,020 nm was used as a radiation source. The laser pulse was divided by the beam splitter into the reference and signal pulses. The reference pulse passed through the optical delay line. The signal pulse was reflected from the sample. Both beams were then focused on the nonlinear BBO crystal, and noncollinear second-harmonic generation was detected by a PMT. The angle of incidence of 67° was chosen to overlap the spectral range of the plasmon resonance with the laser tuning range. The experimental setup allowed for the measurements for both p- and s-polarized incident light. The dependence of the PMT signal on the delay time between pulses is the second order intensity correlation function (CF). The correlation function of the s-polarized pulse is an autocorrelation function since there is no excitation of SPPs and the laser pulse reflects from the sample without any perturbation. Measurements of CFs are performed in the spectral range from 710 to 800 nm.

Figure 2.19 shows the normalized correlation function measured for p- and s-polarized light at the resonant and nonresonant wavelengths of 722 nm

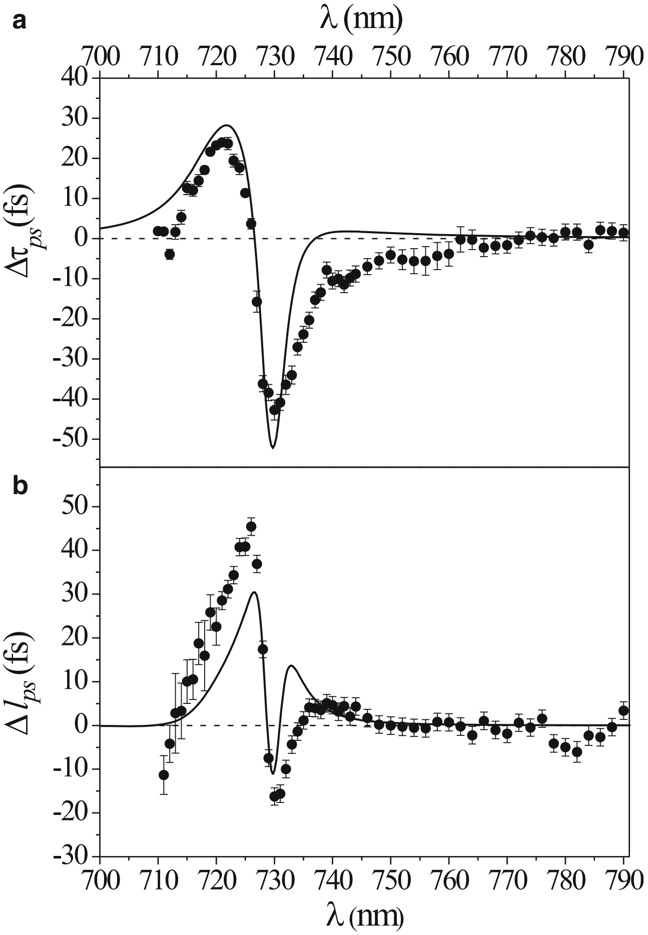


Fig. 2.20 The differential spectrum of the CF peak positions $\varDelta\tau_{\mathrm{ps}}$ (**a**) and the width of CF $\varDelta l_{\mathrm{ps}}$ (**b**). The points stand for the experimental data, and the *solid curves* stand for the numerical calculations [76]

and 780 nm, respectively. The pronounced modification of the CF shape and shift of the CF maximum are observed in the vicinity of the resonance for the p-polarized pulses with respect to the s-polarized ones, while CFs for the p- and s-polarized pulses measured out of the resonance look similar. Figure 2.20a shows spectral dependence of the time shift between CF maxima for the p- and s-polarized pulses. The Fano-type interference gives the effect of both delay and leading of the p-polarized pulse relative to s-polarized pulse, as well as its broadening and narrowing in the vicinity of the resonance. The broadening of the CF measured for the p-polarized pulse and its delay in comparison to the nonresonant pulse is associated with the relaxation of resonantly excited SPPs. In the vicinity of the Fano

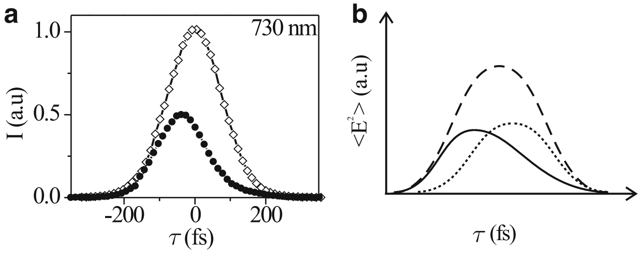


Fig. 2.21 (**a**) The second order intensity correlation function for p-polarized (*filled circles*) and s-polarized (*empty circles*) pulses at $\lambda = 730$ nm. (**b**) The schematic image of the destructive interference between the reflected pulse nonresonant component (*dashed line*) and delayed SPP excitation (*dotted line*). The *solid line* represents the resulting pulse reflected from the sample [76]

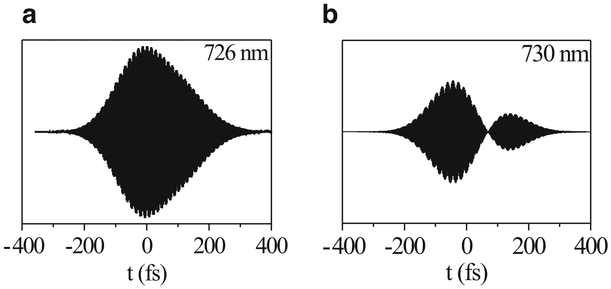


Fig. 2.22 The calculated time-dependent electric field of the pulse reflected from the sample. (**a**) Wavelength $\lambda = 726$ nm corresponds to the maximal CF broadening. (**b**) Wavelength $\lambda = 730$ nm corresponds to the minimum of the Fano resonance [76]

resonance minimum the CF amplitude measured for the p-polarized pulse and, consequently, the intensity of the pulse reflected from the sample is smaller than that of the s-polarized pulse as it is shown in Fig. 2.21a. The asymmetry of the CF shape and the CF maximum shift are clearly seen due to the destructive interference between the nonresonant and resonant reflected components of the p-polarized pulse as a result of an additional phase difference between them (Fig. 2.21b). If the pulse which is nonresonantly reflected from the sample is in phase with the pulse which is resonantly reemitted due to SPP excitation, they interfere constructively leading to the CF maximum shift to the positive time values and to the CF broadening. If the pulses interfere destructively, the CF peak shifts to the negative direction and the CF narrows.

Calculated spectral dependences of the CF shift and the CF width difference are shown by lines in Fig. 2.20. A good agreement with the experiments is seen. The result of the reconstruction of the electric field strength of the pulse reflected from the sample is shown in Fig. 2.22. The pulse is broadened in comparison to the reference pulse due to the SPP excitation 726 nm corresponding to the maximal CF broadening (Fig. 2.22a). Significant changes of the profile are also observed at the

Fano resonance minimum at 730 nm (Fig. 2.22b) due to the destructive interference. To conclude, the control over ultrafast optical response of plasmonic nanostructures can be attained for laser pulses duration comparable to the SPP lifetime.

2.8 Magnetoplasmonics

External magnetic field is prospective for fabrication of actively controlled plasmonic components which can be used in a wide number of applications. However, weak magneto-optical effects stemming from the weak spin-orbit interaction stimulate the search for the ways of enhancement of magneto-optical phenomena. One of the examples of photonic devices with magneto-optical response enhanced by proper nanostructuring is magnetophotonic crystals [77, 78] which are periodically structured magnetic dielectrics with the period comparable to the wavelength of the optical range. Last decade magnetophotonic crystals were in the focus of attention due to unique abilities of light propagation control they provide utilizing nonreciprocity of magneto-optical effects [79, 80]. In such structures significant Faraday angles and magneto-optical Kerr effect (MOKE) values were achieved [80–85]. Spatial localization of light was successfully used for enhancement of magnetic-field-induced nonlinear-optical response of magnetophotonic crystals [79, 81, 86–89]. Another approach for enhancement of magneto-optical response deals with periodically structured magnetic metallic materials, where propagating SPP can be excited due to the phase-matching between wave vectors of incident light and SPP and the vector of reciprocal lattice [90–96]. In analogy to magnetophotonic crystals, such materials can be considered as magnetoplasmonic crystals controlling SPP generation and propagation by periodicity.

Below we discuss briefly how longitudinal and transversal configurations of MOKE can be used for studying the spectral dependence of magneto-optical response enhancement at the Wood’s anomaly of one- and two-dimensional (1D and 2D) magnetoplasmonic crystals based on nanostructured nickel films [92, 94–96]. Resonant excitation of SPPs leads to asymmetrical Fano-type spectral profiles of Kerr rotation and relative changes in reflectivity observed in the longitudinal and transversal magnetic field application, respectively. Such a lineshape was associated with Faraday and Voigt configurations of magnetoplasmon excitation at the surface of magnetoplasmonic crystals.

The atomic force microscopy and scanning electron microscopy images of the magnetoplasmonic crystal samples are shown in Fig. 2.23. The images show good lateral periodicity of both samples. The 1D magnetoplasmonic crystal sample was made by nanoimprint lithography in a 100-nm-thick nickel film forming a diffraction grating with a period of 320 nm and a modulation depth of 50 nm. The 2D magnetoplasmonic crystal consisted of a 2D hexagonal array of nickel nanodiscs with a height of 50 nm, a diameter of 200 nm, and a distance between centers of the neighboring discs of 400 nm.

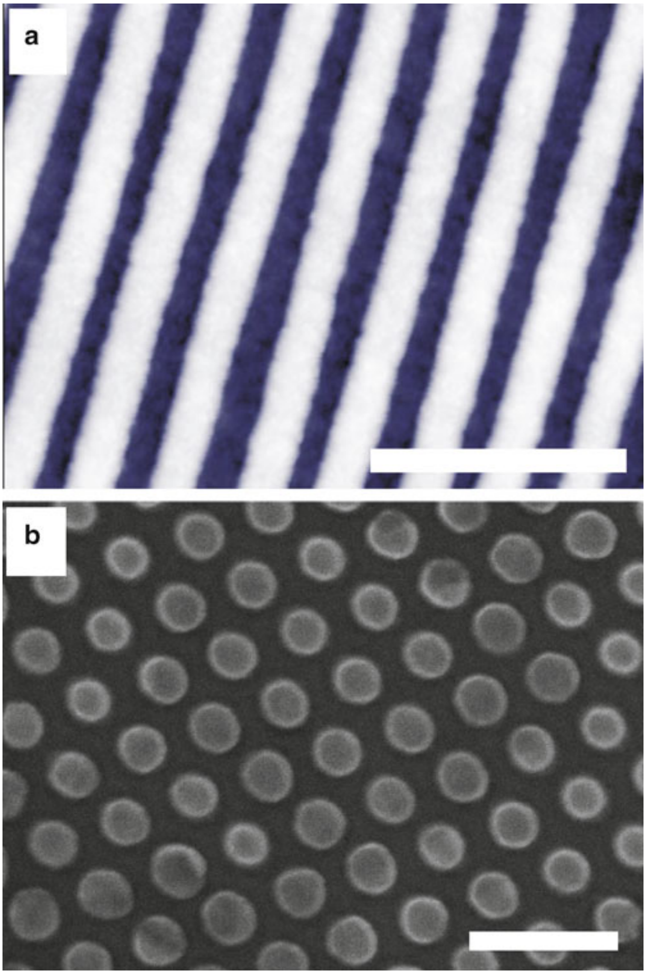


Fig. 2.23 (**a**) An atomic force microscopy image of the 1D magnetoplasmonic crystal sample. (**b**) A scanning electron microscopy image of the 2D magnetoplasmonic crystal sample. Scale bars in both images are equal to 1 μm

Two configurations of the external magnetic field application along the sample surface were used, namely the transversal one as magnetic field is perpendicular to the plane of incidence and the longitudinal one as the magnetic field is directed along the plane of incidence. The value of transversal magneto-optical Kerr effect (TMOKE) was defined as $\delta = (R(\mathbf{M}) - R(-\mathbf{M}))/R_0$, where $R(\mathbf{M})$, $R(-\mathbf{M})$, and R_0 are reflectance with and without the external magnetic field. Longitudinal magneto-optical Kerr effect (LMOKE) was characterized by the Kerr rotation angle of the linearly polarized radiation upon the reflection from the magnetized sample.

The spectral dependences of reflectance and TMOKE measured in the 1D magnetoplasmonic crystal are shown in Fig. 2.24.

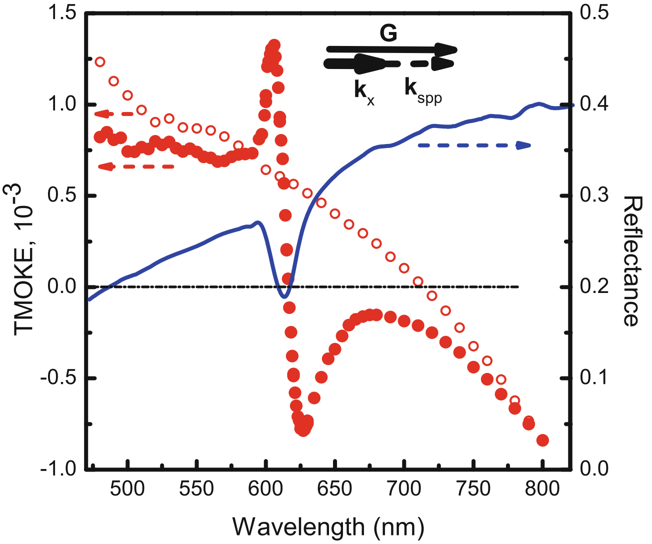


Fig. 2.24 TMOKE (*filled circles*) and reflectance (*curve*) spectra of 1D magnetoplasmonic crystal for the reciprocal vector oriented within the plane of light incidence. *Open circles*: TMOKE spectrum of 1D magnetoplasmonic crystal with the reciprocal vector perpendicular to the incident plane. Angle of incidence is 60°. *Inset*: schematic of the SPP phase-matching conditions with k_x is projection of the light wave vector onto the sample surface and k_{spp} is the SPP wave vector

Reflectance spectrum for p-polarized light measured in specular direction as the reciprocal vector **G** is oriented within the plane of incidence shows the fulfillment of the phase-matching conditions for the excitation of SPPs at the wavelength of 615 nm leading to a resonant dip associated with the Wood's anomaly. The spectral dependence of TMOKE value has a resonance in the vicinity of the Wood's anomaly with an asymmetric lineshape typical for the Fano-type resonance. The reference TMOKE spectrum measured for the sample azimuthal orientation as reciprocal vector is perpendicular to the incident plane providing the spectral behavior similar to that of a plain nickel film. The spectrum decays monotonously with the wavelength increase as it was expected for a nonstructured nickel surface.

The TMOKE is a result of the SPP dispersion curve shift upon the magnetization reversal at the surface of magnetoplasmonic crystals. The reflectivity minimum related to the Wood's anomaly is also spectrally shifted. Since the TMOKE value is proportional to the reflectivity difference for opposite magnetization directions, the spectral dependence of TMOKE has an asymmetric resonance lineshape. In the presence of the external magnetic field, SPPs have properties of magnetoplasmons [97] and TMOKE is related to the Voigt configuration of magnetoplasmons. Surface magnetoplasmon modes in the Voigt configuration are asymmetric with respect to the magnetization direction. The dispersion of SPPs reveals a magnetic-field-induced shift since the wave vector of magnetoplasmonic mode, k_{spp}^{M}, depends

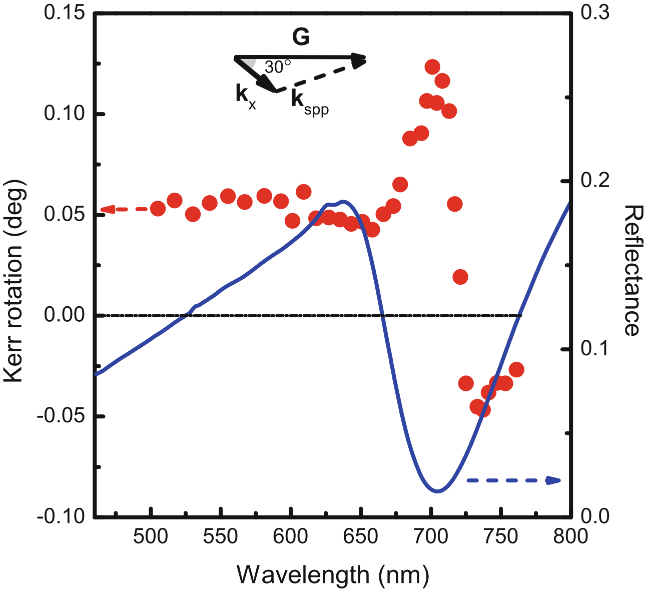


Fig. 2.25 The spectral dependences of the Kerr rotation angle in the longitudinal MOKE configuration (*circles*) and reflectance (*curve*) measured in the 2D magnetoplasmonic crystal. The angle of incidence is 45°. *Inset*: The schematic of the SPP phase-matching condition

on the value of the off-diagonal component of dielectric permittivity, g, in the following form [94, 95, 98]:

$$k_{spp}^{M} = k_{spp}\left(1 \pm \frac{g}{\xi}\right), \tag{2.16}$$

where

$$\xi = \sqrt{\varepsilon_0 \varepsilon_1}\left(1 - \frac{\varepsilon_0^2}{\varepsilon_1^2}\right), \quad k_{spp} = \frac{\omega}{c}\sqrt{\frac{\varepsilon_0 \varepsilon_1}{\varepsilon_0 + \varepsilon_1}} \tag{2.17}$$

ε_0 is the dielectric permittivity of the space outside the magnetoplasmonic crystal and ε_1 is the diagonal component of the nickel dielectric permittivity. The phase matching conditions for the SPP excitation achieved for the opposite magnetization directions are fulfilled for different wavelengths resulting in an asymmetric TMOKE lineshape.

Figure 2.25 shows LMOKE spectrum measured in 2D magnetoplasmonic crystal. The sample was oriented exactly in between two adjacent reciprocal vectors of the hexagonal lattice so that the reciprocal vector **G** is oriented at 30° with respect to the plane of incidence as schematically shown in the inset of Fig. 2.25. The specular reflectance spectra for p-polarized light shows a Wood's anomaly at 700 nm. The Kerr rotation angle θ_K is enhanced up to 0.15° at the wavelength coinciding with the dip in the reflectance spectra and changes the sign upon spectral tuning across the Wood's anomaly.

LMOKE corresponds to the Faraday configuration of magnetoplasmonic modes excitation. In this case plasmonic wave contains both TM and TE components, and SPP can be represented as a coherent superposition of two components with right and left conical polarizations [94, 95]. Within the approximation, which is linearly dependent on g, the dispersion relation for these components is written as follows:

$$k_{spp}^{R(L)} = k_{spp}\left(1 \pm \frac{1}{2}\frac{\varepsilon_0}{(\varepsilon_0 + \varepsilon_1)\varepsilon_1} g\right). \qquad (2.18)$$

The total amplitude of the SPP is determined by the coherent sum of these components, and the SPP polarization is defined by the phase shift between them. The asymmetric lineshape of the Kerr rotation spectra at the Wood's anomaly is attributed to slightly different phase-matching conditions for SPP modes with the wave vectors ${k_{\mathrm{spp}}}^R$ and ${k_{\mathrm{spp}}}^L$.

Summarizing, spectral dependences of transversal and longitudinal magneto-optical Kerr effects measured in one- and two-dimensional magnetoplasmonic crystals possesses a resonant Fano-type enhancement in the angular and wavelength vicinity of the Wood's anomaly, which is associated with the phase-matched excitation of the magnetoplasmon modes excited in Voigt and Faraday configurations, respectively. One-dimensional perturbation of the magnetic metal surface leads to the pronounced resonances in the magneto-optical Kerr effects, while two-dimensional periodicity gives one the possibility for the spectral tuning of the SPP phase-matching conditions by using the superposition of two reciprocal vectors.

2.9 Conclusions

With many problems yet unsolved nanoplasmonics has already proven to be a highly promising area of photonics both from fundamental research and applications point of view. Surface plasmons reveal ways of subwavelength light localization and transport and allow of engineering of novel materials that are capable of efficient control over properties of light by means of optical anisotropy and magneto-optic effects.

Acknowledgements Authors would like to thank A.A. Ezhov, A.V. Chetvertukhin, A.A. Grunin, P.P Vabishchevich, and B.B. Tsema for fruitful discussions.

References

1. V.M. Agranovich, D.L. Mills, *Surface Polaritons - Electromagnetic Waves at Surfaces and Interfaces* (Elsevier Science, New York, 1982)
2. H. Raether, *Surface-Plasmons on Smooth and Rough Surfaces and on Gratings* (Springer, Berlin, 1988)

3. W.L. Barnes, J. Opt. A Pure App. Opt. **8**(4, Sp. Iss. SI), S87 (2006)
4. H.C. van de Hulst, *Light Scattering by Small Particles* (Dover, New York, 1981)
5. K. Kelly, E. Coronado, L. Zhao, G. Schatz, J. Phys. Chem. B **107**(3), 668 (2003)
6. R.W. Wood, Philos. Mag. **4**(19–24), 396 (1902)
7. U. Fano, Phys. Rev. **124**(6), 1866 (1961)
8. A. Drezet, A. Hohenau, D. Koller, A. Stepanov, H. Ditlbacher, B. Steinberger, F.R. Aussenegg, A. Leitner, J.R. Krenn, Mat. Sci. Eng. B **149**(3), 220 (2008)
9. J.A. Veerman, A.M. Otter, L. Kuipers, N.F. van Hulst, Appl. Phys. Lett. **72**, 3115 (1998.)
10. R.J. Moerland, N.F. van Hulst, H. Gersen, L. Kuipers, Opt. Express **13**, 1604 (2005)
11. O. Marti, H. Bielefeldt, B. Hecht, S. Herminghaus, P. Leiderer, J. Mlynek, Opt. Commun. **96**, 225 (1993)
12. M. Balistreri, H. Gersen, J. Korterik, L. Kuipers, N. van Hulst, Science **294**(5544), 1080 (2001)
13. D.K. Gramotnev, S.I. Bozhevolnyi, Nat. Photonics **4**(2), 83 (2010)
14. S. Bozhevolnyi, V. Volkov, E. Devaux, J. Laluet, T. Ebbesen, Nature **440**(7083), 508 (2006)
15. S. Bozhevolnyi, J. Erland, K. Leosson, P. Skovgaard, J. Hvam, Phys. Rev. Lett. **86**(14), 3008 (2001)
16. A. Ezhov, S. Magnitskii, N. Maslova, D. Muzychenko, A. Nikulin, V. Panov, JETP Lett. **82**, 599 (2005)
17. R. Zia, M.L. Brongersma, Nat. Nanotechnol. **2**, 426 (2007)
18. V.M. Shalaev, Nat. Photonics **1**(1), 41 (2007)
19. T.W. Ebbesen, H.J. Lezec, H.F. Ghaemi, T. Thio, P.A. Wolff, Nature **391**, 667 (1998)
20. V.G. Veselago, Sov. Phys. Uspekhi **10**, 509 (1968)
21. R.A. Shelby, D.R. Smith, S. Schultz, Science **292**(5514), 77 (2001)
22. A.N. Grigorenko, A.K. Geim, H.F. Gleeson, Y. Zhang, A.A. Firsov, I.Y. Khrushchev, J. Petrovic, Nature **438**(7066), 335 (2005)
23. V. Shalaev, W. Cai, U. Chettiar, H. Yuan, A. Sarychev, V. Drachev, A. Kildishev, Opt. Lett. **30**(24), 3356 (2005)
24. H.A. Bethe, Phys. Rev. **66**(7-8), 163 (1944)
25. F.J. Garcia-Vidal, F. Lopez-Tejeira, J. Bravo-Abad, L. Martin-Moreno, *Surface Plasmon Nanophotonics* (Springer, New York, 2007)
26. H.F. Ghaemi, T. Thio, D.E. Grupp, T.W. Ebbesen, H.J. Lezec, Phys. Rev. B **58**(11), 6779 (1998)
27. E. Altewischer, C. Genet, M.P. van Exter, J.P. Woerdman, P.F.A. Alkemade, A. van Zuuk, E. W.J.M. van der Drift, Opt. Lett. **30**(1), 90 (2005)
28. J. Porto, F. García-Vidal, J. Pendry, Phys. Rev. Lett. **83**(14), 2845 (1999)
29. A. Barbara, P. Quémerais, E. Bustarret, T. Lopez-Rios, Phys. Rev. B **66**(16), 161403 (2002)
30. Q. Cao, P. Lalanne, Phys. Rev. Lett. **88**(5), 3 (2002)
31. P. Lalanne, J.C. Rodier, J.P. Hugonin, J. Opt. A **7**(8), 422 (2005)
32. Q.J. Wang, J.Q. Li, C.P. Huang, C. Zhang, Y.Y. Zhu, Appl. Phys. Lett. **87**(9), 091105 (2005)
33. R. Gordon, A. Brolo, A. McKinnon, A. Rajora, B. Leathem, K. Kavanagh, Phys. Rev. Lett. **92**(3), 037401 (2004)
34. S.Y. Hsu, K.L. Lee, E.H. Lin, M.C. Lee, P.K. Wei, Appl. Phys. Lett. **95**(1), 013105 (2009)
35. Y. Pang, R. Gordon, Opt. Express **17**(4), 2871 (2009)
36. M.R. Shcherbakov, P.P. Vabishchevich, M.I. Dobynde, T.V. Dolgova, A.S. Sigov, C.M. Wang, D.P. Tsai, A.A. Fedyanin, JETP Lett. **90**, 433 (2009)
37. M.R. Shcherbakov, M.I. Dobynde, T.V. Dolgova, D.P. Tsai, A.A. Fedyanin, Phys. Rev. B. **82**, 193402 (2010)
38. B. Luk'yanchuk, N.I. Zheludev, S.A. Maier, N.J. Halas, P. Nordlander, H. Giessen, C.T. Chong, Nat. Mat. **9**(9), 707 (2010)
39. R. Vogelgesang, A. Dmitriev, Analyst **135**, 1175 (2010)
40. C. Girard, Rep. Progr. Phys. **68**, 1883 (2005)
41. M. Salerno, J. Krenn, A. Hohenau, H. Ditlbacher, G. Schider, A. Leitner, F. Aussenegg, Opt. Commun. **248**(4-6), 543 (2005)

42. N.C. Lindquist, P. Nagpal, A. Lesuffleur, D.J. Norris, S.H. Oh, Nano Lett. **10**, 1369 (2010)
43. Q. Wang, J. Bu, X.C. Yuan, Opt. Express **18**(3), 2662 (2010)
44. G.M. Lerman, A. Yanai, U. Levy, Nano Lett. **9**(5), 2139 (2009)
45. W. Chen, D.C. Abeysinghe, R.L. Nelson, Q. Zhan, Nano Lett. **10**(6), 2075 (2010)
46. D. Bergman, M. Stockman, Phys. Rev. Lett. **90**(2), 027402 (2003)
47. R.F. Oulton, V.J. Sorger, T. Zentgraf, R.M. Ma, C. Gladden, L. Dai, G. Bartal, X. Zhang, Nature **461**(7264), 629 (2009)
48. J.A. Dionne, K. Diest, L.A. Sweatlock, H.A. Atwater, Nano Lett. **9**(2), 897 (2009)
49. E.B. McDaniel, S.C. McClain, J.W.P. Hsu, Appl. Opt. **37**(1), 84 (1998)
50. M.R. Shcherbakov, B.B. Tsema, Y.B. Tsema, A.A. Ezhov, V.I. Panov, D.P. Tsai, A.A. Fedyanin, Physica C **479**, 183 (2012)
51. M. Shcherbakov, B. Tsema, A. Ezhov, V. Panov, A. Fedyanin, JETP Lett. **93**, 720 (2011)
52. H. Gao, J. Henzie, T.W. Odom, Nano Lett. **6**(9), 2104 (2006)
53. L. Yin, V.K. Vlasko-Vlasov, J. Pearson, J.M. Hiller, J. Hua, U. Welp, D.E. Brown, C.W. Kimball, Nano Lett. **5**(7), 1399 (2005)
54. S. Kitson, W. Barnes, J. Sambles, Phys. Rev. Lett. **77**(13), 2670 (1996)
55. I. Radko, T. Sondergaard, S. Bozhevolnyi, Opt. Express **14**(9), 4107 (2006)
56. M.I. Stockman, Nat. Photonics **2**(6), 327 (2008)
57. M.A. Noginov, G. Zhu, A.M. Belgrave, R. Bakker, V.M. Shalaev, E.E. Narimanov, S. Stout, E. Herz, T. Suteewong, U. Wiesner, Nature **460**(7259), 1110 (2009)
58. M. Noginov, G. Zhu, M. Mayy, B. Ritzo, N. Noginova, V. Podolskiy, Phys. Rev. Lett. **101**(22), 226806 (2008)
59. E. Plum, V.A. Fedotov, P. Kuo, D.P. Tsai, N.I. Zheludev, Opt. Express **17**(10), 8548 (2009)
60. J. Seidel, S. Grafström, L. Eng, Phys. Rev. Lett. **94**(17), 177401 (2005)
61. N.I. Zheludev, S.L. Prosvirnin, N. Papasimakis, V.A. Fedotov, Nat. Photonics **2**(6), 351 (2008)
62. J.N. Caspers, N. Rotenberg, H.M. van Driel, Opt. Express **18**(19), 19761 (2010)
63. S. Herminghaus, P. Leiderer, Appl. Phys. A Mater. Sci. Process. **51**(4), 350 (1990)
64. R.H.M. Groeneveld, R. Sprik, A. Lagendijk, Phys. Rev. Lett. **64**, 784 (1990)
65. C. Sönnichsen, T. Franzl, T. Wilk, G. von Plessen, J. Feldmann, Phys. Rev. Lett. **88**(7), 077402 (2002)
66. D. Kim, S. Hohng, V. Malyarchuk, Y. Yoon, Y. Ahn, K. Yee, J. Park, J. Kim, Q. Park, C. Lienau, Phys. Rev. Lett. **91**(14), 143901 (2003)
67. A.S. Vengurlekar, A.V. Gopal, T. Ishihara, Appl. Phys. Lett. **89**(18), 181927 (2006)
68. M.R. Shcherbakov, P.P. Vabishchevich, V.V. Komarova, T.V. Dolgova, A.A. Fedyanin, Phys. Rev. Lett. **108**, 253903 (2012)
69. T. Zentgraf, A. Christ, J. Kuhl, H. Giessen, Phys. Rev. Lett. **93**(24), 243901 (2004)
70. A. Dogariu, T. Thio, L.J. Wang, T.W. Ebbesen, H.J. Lezec, Opt. Lett. **26**(7), 450 (2001)
71. C. Ropers, D. Park, G. Stibenz, G. Steinmeyer, J. Kim, D. Kim, C. Lienau, Phys. Rev. Lett. **94**(11), 113901 (2005)
72. H. Su, Z. Hang, Z. Marcet, H. Chan, C. Chan, K. Wong, Phys. Rev. B **83**(24), 245449 (2011)
73. C. Ropers, G. Stibenz, G. Steinmeyer, R. Müller, D. Park, K. Lee, J. Kihm, J. Kim, Q. Park, D. Kim, C. Lienau, Appl. Phys. B **84**(1-2), 183 (2006)
74. T. Utikal, T. Zentgraf, J. Kuhl, H. Giessen, Phys. Rev. B **76**(24), 245107 (2007)
75. J. Li, H. Iu, D.Y. Lei, J.T.K. Wan, J.B. Xu, H.P. Ho, M.Y. Waye, H.C. Ong, Appl. Phys. Lett. **94**(18), 183112 (2009)
76. P.P. Vabishchevich, V.O. Bessonov, F.Y. Sychev, M.R. Shcherbakov, T.V. Dolgova, A.A. Fedyanin, JETP Lett. **92**(9), 575 (2011)
77. M. Inoue, K. Arai, T. Fujii, M. Abe, J. Appl. Phys. **85**, 5768 (1999)
78. M. Inoue, R. Fujikawa, A. Baryshev, A. Khanikaev, P. Lim, H. Uchida, O.A. Aktsipetrov, A. A. Fedyanin, T.V. Murzina, A.B.. Granovsky, J. Phys. D Appl. Phys. **39**, R151 (2006)
79. A.A. Fedyanin, T. Yoshida, K. Nishimura, G. Marowsky, M. Inoue, O.A. Aktsipetrov, JETP Lett. **76**, 527 (2002)
80. S. Kahl, A. Grishin, Appl. Phys. Lett. **84**, 1438 (2004)

81. A.A. Fedyanin, T. Yoshida, K. Nishimura, G. Marowsky, M. Inoue, O.A. Aktsipetrov, J. Magn. Magn. Mater. **258–259**, 96 (2003)
82. A.A. Fedyanin, O.A. Aktsipetrov, D. Kobayashi, K. Nishimura, H. Uchida, M. Inoue, J. Magn. Magn. Mater. **282**, 256 (2004)
83. R. Li, M. Levy, Appl. Phys. Lett. **86**, 251102 (2005)
84. A.G. Zhdanov, A.V. Fedyanin, O.A. Aktsipetrov, D. Kobayashi, H. Uchida, M. Inoue, J. Magn. Magn. Mater. **300**, e253 (2006)
85. A.B.. Khanikaev, A.V. Baryshev, P.B. Lim, H. Uchida, M. Inoue, A.G. Zhdanov, A.A. Fedyanin, A.I. Maydykovskiy, O.A. Aktsipetrov, Phys. Rev. B **78**, 193102 (2008)
86. T.V. Dolgova, A.A. Fedyanin, O.A. Aktsipetrov, K. Nishimura, H. Uchida, M. Inoue, J. Appl. Phys. **95**, 7330 (2004)
87. T.V. Murzina, R. Kapra, T.V. Dolgova, A.A. Fedyanin, O.A. Aktsipetrov, K. Nishimura, H. Uchida, M. Inoue, Phys. Rev. B **70**, 012407 (2004)
88. O.A. Aktsipetrov, T.V. Dolgova, A.A. Fedyanin, R. Kapra, T.V. Murzina, K. Nishimura, H. Uchida, M. Inoue, Laser Phys. **14**, 685 (2004)
89. O.A. Aktsipetrov, T.V. Dolgova, A.A. Fedyanin, T.V. Murzina, M. Inoue, K. Nishimura, H. Uchida, J. Opt. Soc. Am. B **22**, 176 (2005)
90. A.B.. Khanikaev, A.V. Baryshev, A.A. Fedyanin, A.B.. Granovsky, M. Inoue, Opt. Express **15**, 6612 (2007)
91. C. Clavero, K. Yang, J. Skuza, R. Lukaszew, Opt. Lett. **35**, 1557 (2010)
92. A.A. Grunin, A.G. Zhdanov, A.A. Ezhov, E.A. Ganshina, A.A. Fedyanin, Appl. Phys. Lett. **97**, 261908 (2010)
93. V.I. Belotelov, I.A. Akimov, M. Pohl, V.A. Kotov, S. Kasture, A.S. Vengurlekar, A. Gopal, D. Yakovlev, A. Zvezdin, M. Bayer, Nat. Nanotechnol. **6**, 370 (2011)
94. A.V. Chetvertukhin, A.V. Baryshev, H. Uchida, M. Inoue, A.A. Fedyanin, J. Appl. Phys. **111**, 07A946 (2012)
95. A.V. Chetvertukhin, A.A. Grunin, A.V. Baryshev, T.V Dolgova, H. Uchida, M. Inoue, A.A. Fedyanin, J. Magn. Magn. Mater. **324**, 3516 (2012)
96. A.A. Grunin, N. Sapoletova, K. Napolskii, A. Eliseev, A.A. Fedyanin, J. Appl. Phys. **111**, 07A948 (2012)
97. R.F. Wallis, in *Surface Magnetoplasmons on Semiconductors*, ed. by A.D. Boardman. Electromagnetic Surface Modes (Wiley, New York, 1982)
98. M. Kushwaha, P. Halevi, Phys. Rev. B **36**, 5960 (1987)

Chapter 3
Nanoscale Photovoltaics and the Terawatt Challenge

Stephen M. Goodnick, Nikolai Faleev, and Christiana Honsberg

Abstract Achieving a sustainable energy system providing terawatts (TWs) of electricity is one of the defining challenges of the coming decades. Photovoltaic technology provides the most likely path to realizing TW scale conversion of solar energy in the future and has been on a nearly 40% growth curve over the past two decades. In order to maintain this rapid level of growth, innovations in cell design and conversion efficiency are needed that are compatible with existing technology and can lead to improved performance and lower cost. Nanotechnology offers a number of advantages to realizing such innovation, by providing new materials and the implementation of advanced concepts that circumvent the current physical limits on efficiency. This chapter reviews several of the promising applications of nanotechnology to photovoltaic technologies and their prospects for the future.

3.1 Terawatt Challenge in Photovoltaics

3.1.1 Introduction to Photovoltaics

Achieving a sustainable energy system is one of the defining challenges of the coming decades. The importance and difficulty of developing an energy source which can replace the existing electrical infrastructure is often termed the "Terawatt Challenge," referring to the fact that to supply the energy demands of the globe, truly massive amounts of electricity—terawatts (TW)—are needed [1, 2]. Since the term received broad public airing in 2005, addressing the TW challenge has been given additional impetus and become time critical due to a convergence of multiple issues, including global warming; the scarcity, uneven geographical

S.M. Goodnick (✉) • N. Faleev • C. Honsberg
School of Electrical, Computer and Energy Engineering, Arizona State University,
P.O. Box 875706, Tempe, AZ 85287-5706, USA
e-mail: stephen.goodnick@asu.edu

A. Korkin and D.J. Lockwood (eds.), *Nanoscale Applications for Information and Energy Systems*, Nanostructure Science and Technology,
DOI 10.1007/978-1-4614-5016-0_3, © Springer Science+Business Media New York 2013

distribution and cost of conventional resources, particularly oil[1]; and other factors such as aging existing infrastructure, the desirability of local electricity production, or the creation of jobs. At the same time as meeting the TW challenge has become more important, the continuing rapid expansion of renewable technologies has simultaneously offered a solution to the seemingly intractable problem. In the seven years since the term was coined, the installed photovoltaic capacity has quadrupled, and Germany, Spain, and Japan's yearly PV installations meet these countries' annual increase in electricity demand. Yet even in these countries, the overall fraction of the electricity demand met by PV is only a few percent. In order for PV to make a larger contribution, both in the US and worldwide, it must continue its rapid growth and expansion.

Photovoltaic energy conversion is the newest of the energy conversion mechanisms which are suggested for large-scale electricity generation. While the effect was first recorded in 1848 by Bequerel with selenium diodes, its understanding and development depended on the theoretical underpinnings provided by quantum mechanics and solid state theory, as well as on the enormous growth and infrastructure associated with the semiconductor industry. The first practical solar cell was demonstrated in 1954 at Bell Laboratories. Impressively, the first module was installed a few years later in Georgia. Perhaps even more so than the transistor, the first solar cells must have seemed astonishing—batteries at the time were relatively large, so something so small that could power a radio was truly an innovation.

Since that time, the transistor and many other semiconductor devices have traced a trajectory of explosive growth and impact, dramatically improving performance, even exceeding assumed fundamental performance limits. Today, photovoltaics has seen the beginning of similarly explosive growth, and since the mid-2000s has used more silicon than the semiconductor industry. As illustrated in Fig. 3.1, the learning curve for photovoltaic modules, along with that for balance of system (BOS) components (the costs not directly associated with the cell and module itself) shows a rapid, constant learning rate over six orders of magnitude of cumulative production. Despite this fact, photovoltaics is just at the beginning of further growth, and to impact on the overall energy picture, it must grow by another two orders of magnitude. Importantly, as shown in Fig. 3.1, if photovoltaics maintains its historical growth rates, within a decade it will make large and substantial contributions to electricity generation, with yearly world electricity generation equal to the US total electricity demand and with the new installation installed from then on exceeding the yearly increases in world electricity demand.

Despite the impressive growth and increasing impact of photovoltaics, as well as the continually increasing efficiency, the performance of photovoltaics still lags far behind that predicted by thermodynamics. In contrast to both a semiconductor device technology or an electrical generating technology, photovoltaics has

[1] Oil is not used for large-scale electricity production, but is coupled through suggestions of shifting transport demand to either natural gas derivatives (which impacts peaking power for electricity) or directly though electric hybrids.

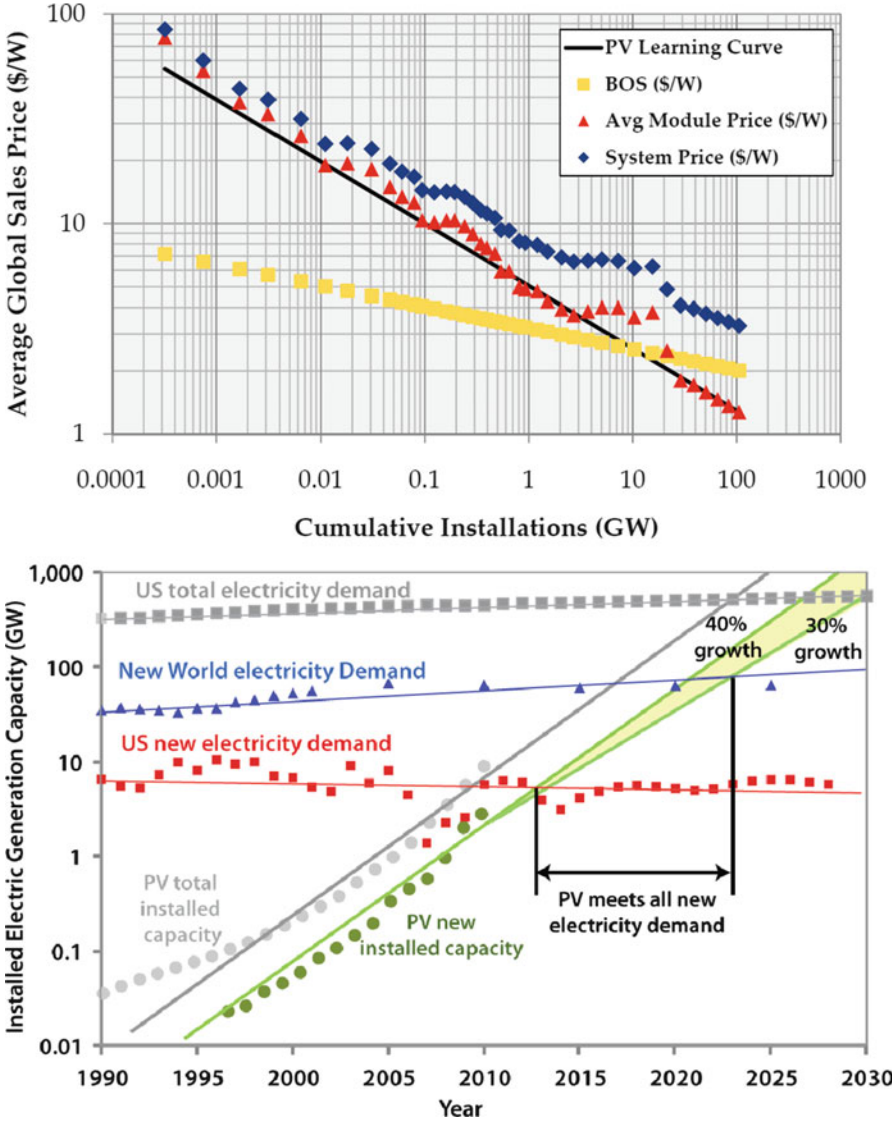


Fig. 3.1 Photovoltaic cumulative production and annually added photovoltaics capacity compared to new US and world electricity demand and the total US electricity demand. The two curves for new installed PV capacity represent continued growth and 40 % compound annual growth rate (CAGR) and 30 % CAGR

achieved less than half of its thermodynamic maximum. On an individual material basis, it does much better, achieving about 75–80% of its material-specific thermodynamic limits, which still leaves substantial room for improvement compared to other devices or electricity systems. Thus, one of the central questions in photovoltaics is why have efficiencies not reached similar levels of optimization as in other technologies. At first, this may seem like an important problem only for researchers. However, both the experience curve and the costing equations for photovoltaics show that achieving high efficiency at low cost is a critical goal for photovoltaics. The answer to the question of why existing semiconductor approaches have not reached thermodynamics energy conversion limits lies in the fact that the principal technologies of conventional semiconductor devices based on pn junctions and metal oxide semiconductor interfaces are poorly suited to implementing the optimum device structure thermodynamically. New physical mechanisms and material properties enabled by nanotechnology suggest approaches to realizing thermodynamic efficiency limits, but require development and exploitation of effects which are not dominant or even nonexistent in conventional approaches. In the following, we briefly review why efficiency is important for photovoltaics, the thermodynamic efficiency limits and implications for photovoltaics, and how nanostructures can improve existing devices and allow devices near the thermodynamic maximum.

A simple equation for the cost of electricity (COE) in \$/kWh for photovoltaics is given by:

$$\mathrm{COE} = \frac{(\$/m^{2}{}_{\mathrm{PV}} + \$/m^{2}_{\mathrm{BOS}})}{\eta \times S} \times \text{Financing Cost} \tag{3.1}$$

where η is the solar cell efficiency, S is the annual incident solar insolation, $\$/\mathrm{m}^2$ PV is the per area cost of the photovoltaics, and $\$/\mathrm{m}^2$ BOS is the cost of balance of system (BOS) components, which as mentioned earlier includes nonpower producing elements, including land, wiring, and power conditioning. Equation (3.1) shows that as the cost of photovoltaics decreases, the BOS costs become an increasing portion of the overall costs. Since many of the BOS or nonpower producing components (glass, land, wiring, etc.) are less subject to large cost reductions, reducing BOS costs by other approaches becomes important. Increasing the efficiency drives costs down in multiple ways. Efficiency not only directly reduces cost by appearing in the denominator of (3.1), a higher efficiency system will have a lower area, and hence BOS costs (the second term in the numerator) are also decreased. Figure 3.2 plots the above equation, which shows that to achieve low eventual system costs, higher efficiencies are necessary.

3.1.2 Limits of Conversion Efficiency

Photovoltaics, compared to other semiconductor devices, have the unusual advantage that the performance (efficiency) can be calculated independent of material assumptions from thermodynamic considerations. One of the first papers to do so is Shockley and Queisser's 1961 paper [3] based on an idealized description of a solar

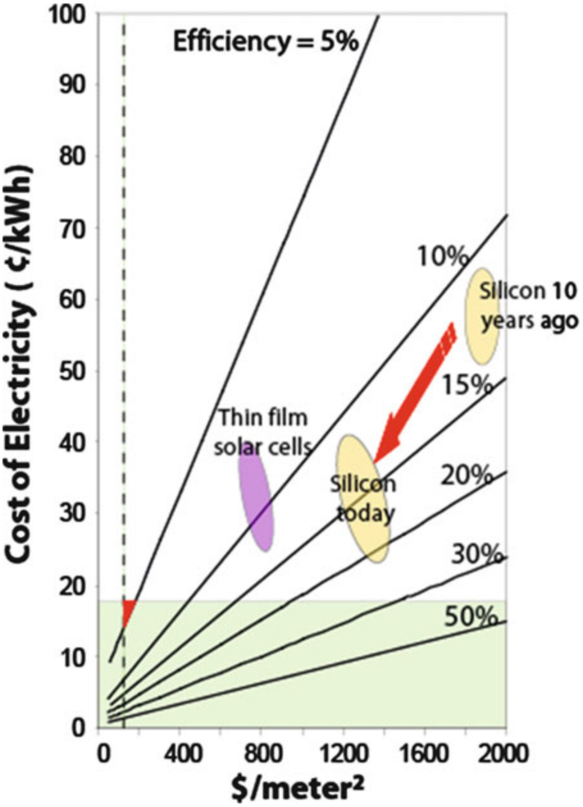


Fig. 3.2 Levelized cost of electricity (LCOE) versus system cost for different conversion efficiencies demonstrating the impact of higher efficiency on the overall cost of electricity

converter which includes no details of the cell structure itself, rather it assumes complete collection of available photogenerated carriers with the following basic assumptions (1) radiative recombination only; (2) one band gap; (3) absorption across the band gap in which one photon generates one electron–hole pair; (4) one associated quasi-Fermi level separation with each band gap; (5) constant temperature in which the carrier temperature is equal to the lattice and ambient temperature; and (6) steady state, which is close to equilibrium. The result of this type of calculation, referred to as *detailed balance*, is illustrated in Fig. 3.3, where the black curve is the calculated efficiency versus bandgap for an AM1.5 solar spectrum (roughly corresponding to terrestrial solar spectrum). The maximum efficiency without concentration is around 33.7% corresponding to maxima around 1.1 and 1.4 eV (the former value being close to the Si bandgap). The principal losses are due to the loss of photons with energy below the bandgap and loss of the excess energy of the photon above the bandgap in terms of energy relaxation of photoexcited carriers back to the band edges. The trade-off between these two loss mechanisms leads to the maximum efficiency as shown in Fig. 3.3.

In their paper, they also examined the effects of nonradiative recombination, and the concept of using multiple bandgaps (tandems) to circumvent the single gap limit

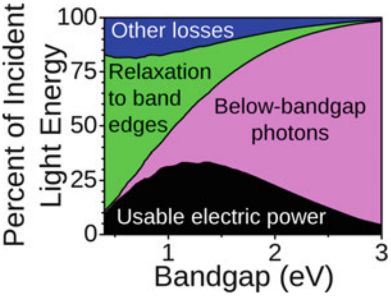


Fig. 3.3 Breakdown of the various loss mechanisms contributing to the single bandgap Shockley–Queisser efficiency limit. The *black shaded region* represents the calculated AM1.5 solar spectrum conversion efficiency as a function of bandgap, while the other *shaded regions* represent the losses (source: Wikipedia Commons)

shown earlier had already been introduced as early as 1955. The next 30 years in photovoltaics focused largely on increasing the efficiency of single and multiple band gap solar cells, and in the mid-1990s silicon had achieved an efficiency of 25% (a "champion result which still stands"). Tandem solar cells also saw efficiency increases, although here the efficiency of champion solar cells is still increasing [4].

During this time, detailed balance and thermodynamic analyses continued to progress, including tandems, photon recycling, and other effects into the analysis. However, with only a small number of exceptions (most notably Nozik and Ross in 1982 [5]), the remainder of the assumptions of Shockley and Queisser were not examined. In 1990, Barnham and Duggan [6] made the suggestion that a quantum well solar cell could achieve the voltage of the larger band gap, but the current of the smaller, and S. Kolodinski et al. measured quantum efficiencies [7]. These papers sparked a re-examination of thermodynamic and detailed balance limits, as well as multiple suggestions of approaches to improve efficiencies above a single junction limit.

Photovoltaics energy conversion is distinct from other electricity generation mechanisms in that it directly converts sunlight into electricity. All other electricity generation mechanisms ultimately rely on a generator, converting mechanical energy into electrical energy. Most of the electricity generation approaches (with the exception of wind and hydropower) are thermodynamically similar in that they convert a heat source into electrical energy. The thermodynamics of such approaches are understood as a heat engine—an energy source heats a reservoir, and energy is extracted from the reservoir. The thermodynamics of most solar cells are a "quantum converter," in which an electron(s) interacts with a photon (or photons). Thermodynamically, photovoltaics is most similar to a Stefan–Boltzmann engine. The energy from the photon must be converted to another state or it is lost—unlike a heat engine where the excess heat may be retained by the thermal reservoir. While the maximum thermodynamic efficiency under maximum concentration and completely optimum assumptions is the same

Table 3.1 Comparison of photothermal and photovoltaic approaches as a function of the number of band gaps and concentration (C)

C	Number of band gaps	Photo-thermal	Photo-voltaic
1	1	53.6	31.0
	2	60.9	42.9
	3	63.3	49.3
	...	...	...
	∞	68.2	68.2
100	1	67.0	35.2
	2	71.7	48.4
	3	73.2	55.6
	...	...	...
	∞	76.2	76.2
46,300	1	85.4	40.8
	2	86.1	55.7
	3	86.3	63.9
	...	...	...
	∞	86.8	86.8

between a quantum converter and a thermal converter, their thermodynamic efficiency under noninfinite concentration or noninfinite number of “band gaps” varies considerably. This fact is presented in Table 3.1, which shows that a thermal converter (e.g. a power plant) is very dependent on concentration (i.e., the operating temperature of the converter). This relationship is well known and understood for thermal power plants, with practical and material limits giving a roughly 30% efficiency for these electricity generating schemes, regardless of whether the heat source is nuclear, coal, or solar thermal. However, unlike the case of power plants, in a solar converter where the carrier temperature drives the thermal engine, the carrier temperature can be substantially elevated under solar concentration, providing increased efficiency, as seen in Table 3.1. In conventional photovoltaic approaches, because they are a quantum converter, the photon energy above the band gap is lost. Concentration makes a small difference in efficiency, largely because the quasi-Fermi levels move closer to the band gap, where the lost energy above the band gap is the largest driving factor, and can only be overcome by using a stack of band gaps (called tandem solar cells). The efficiency as a function of band gap and concentration is shown in Fig. 3.4.

The re-examination of thermodynamics gave rise to the realization that the assumptions 3–6 in the detailed balance examination are not thermodynamically inherent, and that approaches that can overcome these assumptions will give rise to higher efficiencies. A summary of the key approaches that are used to overcome these assumptions is given in Table 3.2. Of these, the ones with the most experimental and theoretical investigation are intermediate band solar cells, multiple exciton solar cells, and hot carrier solar cells. These are discussed in Sect. 3.3.

The above discussion highlights why existing solar cells have not reached their overall thermodynamic efficiency limits, even though they have reached their material-imposed efficiency limits. A p–n junction, even a stack of p–n junctions, is not the thermodynamic ideal—for example, a stack of 8 solar cells or more is

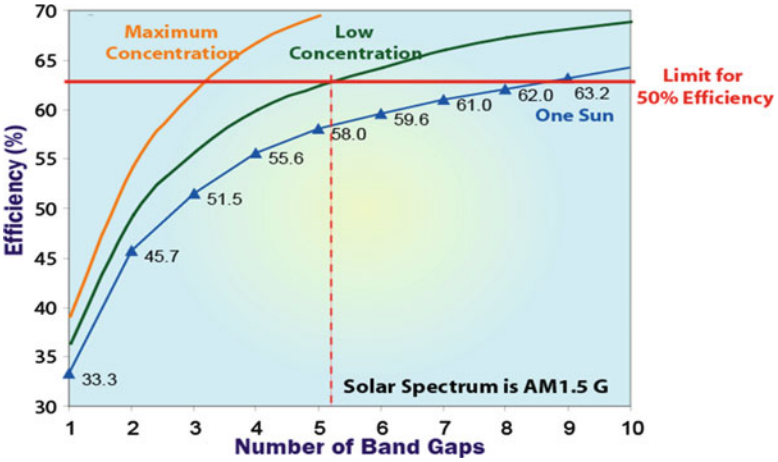


Fig. 3.4 Calculated detailed balance (thermodynamic) efficiency of an ideal solar converter as a function of the number of bandgaps in a tandem cell configuration, for three different concentrations of AM1.5 spectrum sunlight

Table 3.2 Advanced concept solar cell approaches

Assumption in Shockley–Queisser	Approach which circumvents assumption	Examples
Input is solar spectrum	*Multiple spectrum solar cells*: transform the input spectrum to one with same energy but narrower wavelength range	Up/down conversion Thermophotonics
1 photon = 1 electron–hole pair	*Multiple absorption path solar cells*: any absorption path in which one photon ≠ one electron–hole pair	Impact ionization, MEG, two-photon absorption
One quasi-Fermi level separation	*Multiple energy level solar cells*: Existence of multiple meta-stable light-generated carrier populations within a single device	Intermediate band quantum well solar cells
Constant temperature = cell temperature = carrier temperature	*Multiple temperature solar cells*: Any device in which energy is extracted from a difference in carrier or lattice temperatures	Hot carrier solar cells
Steady state (≈ equilibrium)	*AC solar cells*: Rectification of electromagnetic wave	Rectenna solar cells

necessary to realize an efficiency approaching the thermodynamic limit. However, thermodynamics indicates that there are structures or physical mechanisms which reach thermodynamic limits—for example, a quantum/thermal hybrid thermodynamic reaches efficiencies of 68% under one sun—close to the thermodynamic maximum of 70%. While the approach to implementing such structures in not clear with conventional semiconductor materials and approaches, it highlights that new physical mechanisms possible in nanostructured materials are key to realizing ultra-high efficiency solar cells.

3.1.3 Existing Solar Cell Technology

The dominant commercial solar cell technology consists of silicon solar cells, comprising over 80% of the market. The most common silicon solar cell is a front junction, screen printed device, consisting of a diffused junction, a silicon nitride surface passivation, silver screen-printed metal front contacts which are fired through the silicon nitride and aluminum/silver rear contacts. The record silicon solar cell is 25%, and commercial modules typically are 14% for multicrystalline and 17% for monocrystalline, with a few commercial technologies having efficiencies close to the record cell results. Thin film solar cells, mostly CdTe but also CIGS (CuInGaSe2) provide a lower fabrication and lower $/W cost, but with lower efficiency. Overall, the two technologies have similar cost of electricity.

A substantially different technology and approach is to make ultra-high efficiency solar cells by growing multiple solar cells in a single solar cells stack. The high cost of these devices is compensated for by using them in an optical concentrator (where the light intensity is 200× to 400× higher than typical sunlight), such that only very small areas are needed. The optical systems must track the sun, and these large systems are suited primarily for utility scale applications. The central issue in such devices is the availability of materials which are lattice matched or nearly so, and also have ideal band gaps. The lack of materials with 1.0 eV lattice matched to GaAs limits the ability to realize substantial efficiency increases, although metamorphic or dilute nitride (InGaAsN) have demonstrated champion efficiencies.

3.1.4 Advanced Concept Solar Cells

As illustrated in Fig. 3.3, the Shockley–Queisser limit is a consequence of the fact that photons below the bandgap of the absorber are not collected, while each absorbed photon above contributes only the energy of the of electron–hole pair at the bandgap, independent of the photon energy. So-called *third generation photovoltaics* refers to solar cell technology in which advanced concepts usually based on nanotechnology are used to circumvent the single gap limit and help drive cost down through improvement in efficiency [8]. Figure 3.5 shows a slightly different representation of Fig. 3.2, illustrating the expected market for new concept solar cells in terms of the cost of electricity. The advantage of higher efficiency allows for higher module cost, and as mentioned earlier, may lead to significant reductions in the BOS costs as well, reducing the overall cost of electricity.

There are a number of paths to approaching thermodynamic conversion efficiencies (~85%) rather than the single gap Shockley–Queisser limit. Table 3.2 gives a summary of different approaches, which go beyond the assumptions inherent in the single band limit. One of the limitations in efficiency implicit in Fig. 3.3 is the fact that the solar spectrum is a broadband source, which leads to the trade-offs in various energy loss mechanisms and the difficulty in optimizing the performance for

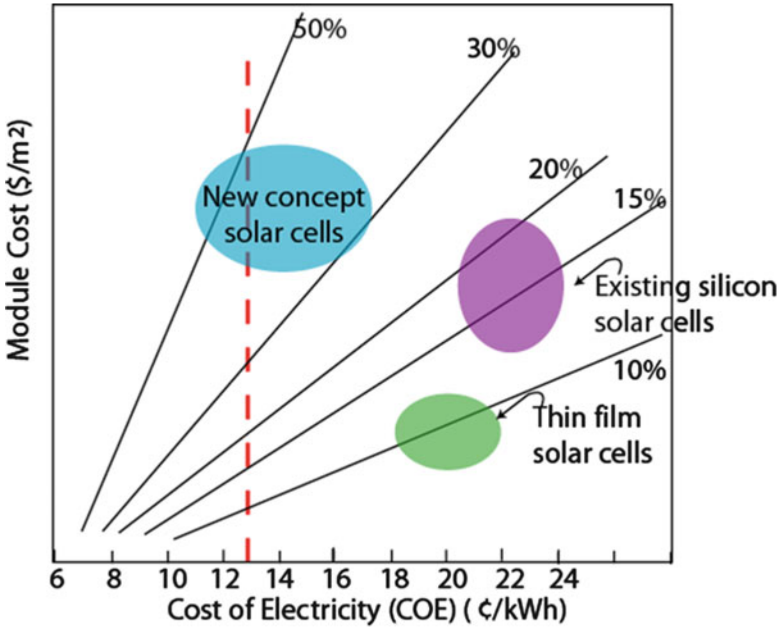


Fig. 3.5 Module cost as a function of levelized cost of electricity (COE) for different conversion efficiencies showing the advantage of high efficiency solar cells in terms of reduced COE

a single gap. If the solar spectrum can be transformed to a narrow spectrum, higher efficiency is possible. Schemes for this include up/down conversion of the solar spectrum through phosphors or two-photon absorption/emission. Conversion of the solar spectrum to thermal energy, and then photovoltaic recapture of the blackbody radiation at a lower temperature (thermophotonics) effectively narrows the solar spectrum as well, and theoretically leads to much higher conversion efficiency.

Another route to exceeding the single gap limit is to generate multiple electron–hole pairs from a single photon through the creation of secondary carriers (band-to-band impact ionization or multiexciton generation) [9]. Particularly promising are nanocrystalline materials, where the reduced dimensionality of the system suppresses competing energy relaxation mechanisms and reduces the threshold for generation of secondary carriers as discussed in Sect. 3.3.2 [10]. The fact that quantum efficiencies greater than unity may be possible, consequently lead to theoretical efficiencies greater than the Shockley–Queisser limit.

Another way to circumvent the single gap limit is to have multiple energy level solar cells. Multijunction or tandem solar cells were already discussed in this context in Sect. 3.1.2 and have shown the highest efficiencies of any solar cell technology [11]. Such cells are grown using epitaxial material growth technology discussed in the next section, which is generally quite expensive compared to conventional Si solar cell production. Hence such cells are targeted for concentrating photovoltaic (CPV) systems, in which high concentration is used to focus sunlight on a small diameter cell area.

An alternate approach to having multiple junctions is to introduce multiple levels within the same material, which provide multiple paths for photon absorption, but collect carriers at the primary bandgap of the host material. Luque and Marti introduced the concept of an intermediate band (IB) solar cell to realize such a structure, and overcome the Shockley–Queisser limit [12], although as mentioned earlier, similar concepts had been suggested for quantum well solar cells. An intermediate level in the bandgap is introduced through, for example, self-assembled quantum dots, which allow low energy photons to excite electron–hole pairs through multiphoton absorption, below the gap of the principal absorber. Such approaches and their implementation with nanostructured systems are discussed in more detail in Sect. 3.3.3.

Finally, one can relax the assumption of constant temperature throughout and allow for nonequilibrium distributions of carriers at different effective energies. Ross and Nozik proposed the concept of hot carrier solar cells [13] 25 years ago as a means to circumvent the limitations imposed by the Shockley–Queisser limit in terms of both the lost of excess kinetic energy and the loss of long wavelength photons. The concept was extended further by Würfel and coworkers considered the effect of impact ionization and secondary carrier generation on the ultimate efficiency of this concept [14, 15]. In this concept, electrons and holes are not collected at the band edges (which limits to the output voltage to the bandgap), rather they are collected through energy selective contacts above and below the conduction and valence band edges, respectively. The absorber material (generally a narrow gap material) suppresses energy loss, so that hot carriers can reach sufficient energy to escape through the energy selective contacts. This concept is discussed later in Sect. 3.3.4.

3.2 Nanostructured Materials and Devices

Nanotechnology, by its name, refers to technology at small scales, literally at nanometer scale dimensions (10^{-9} m). Somewhat arbitrarily, we define nanometer scale to characteristic feature sizes on the order of 100 nm or less in terms of the separation of the micro and nano-worlds. The fact that almost all such structures contain nanoscale features in one form or another has led to "nanotechnology" being regarded as a somewhat broad umbrella encompassing a host of scientific and engineering disciplines.

The nanotechnology "revolution" has been enabled by remarkable advances in atomic scale probes and nanofabrication tools. Structures and images at the atomic scale have been made possible by the invention of the scanning tunneling microscope (STM), and the associated atomic force microscope (AFM), for which Gerd Binning and Heinrich Rohrer from IBM Research Laboratory were awarded the Nobel Prize in 1986 [1]. Such scanning probe microscopy (SPM) techniques allow atomic scale resolution imaging of atomic positions, spectroscopic features, and positioning of atoms on a surface. Top down nanofabrication techniques such as electron-beam, ion-beam, and deep ultra-violet (UV) lithography allow the patterning of features down to tens of nanometers, and AFM techniques can be used to actually position atoms literally with atomic precision.

Concurrently, there have been significant advances in the “bottom up” synthesis and control of self-assembled materials such as nanoparticles, nanowires, molecular wires, and novel states of carbon such as fullerenes, graphene, carbon nanotubes (CNTs), and composites thereof. These advances have led to an explosion of scientific breakthroughs in studying the unique electronic/optical/mechanical properties of these new classes of materials.

At the same time, such nanostructured materials are emerging as new and improved materials for structural components as well as coatings, insulators, and conductors. Self-assembled materials are also being commercially pursued for potential application as components of electronic devices and circuits. Many energy conversion technologies are also benefiting from the ability to synthesize new nanostructured materials to improve performance or replace costly materials with relatively inexpensive alternatives.

3.2.1 Nanomaterials

Nanomaterials usually refer to materials that have structural features on the nanoscale, and in particular their properties stem from these nanoscale dimensions. Such nanomaterials may include quantum wells, nanoparticles, nanopowders, nanoshells, nanowires, nanorods, nanotubes such as CNTs, nanomembranes and nanocoatings, or combinations of these to form nanocomposites. An important feature of nanomaterials for energy applications compared to their bulk counterparts is that the surface-to-volume ratio is greatly enhanced, resulting in fundamental changes in the chemical, electronic, mechanical, and optical properties, in essence creating a new material. Such changes are a result of the different energies associated with surfaces compared to the bulk. This may result in complete changes in the way materials may behave, in terms of their catalytic properties, their chemical bonding, strength, etc. Another effect is the so-called quantum size effect, which like the simple particle in a box, quantizes the motion of electrons in a solid, meaning the allowed energies can only assume certain discrete values. This generally changes the electrical and optical properties of materials. For example, nanoparticles show a blue shift in their absorption spectrum to high frequency due to quantum confinement effects.

3.2.1.1 Quantum Wells and Superlattices

One of the first truly nanoscale fabrication technologies was the development of precision epitaxial material growth techniques such as molecular beam epitaxy (MBE) [16] and metal organic chemical vapor deposition (MOCVD), through which high-quality, lattice-matched heterojunction (junction between two dissimilar materials) semiconductor layered systems could be realized, with atomic precision in the interface quality. A sandwich composed of a narrower bandgap material clad with larger bandgap materials of atomic dimensions is referred to as a quantum well (QW), and when many of these are grown sequentially, they are referred to as a multiquantum well (MQW) system. These systems exhibit strong quantum confinement effects due to the low density of defects at the interface of lattice-matched materials such as GaAs and $Al_xGa_{1-x}As$. If the thickness of the barriers separating large and small bandgap materials is reduced so that the electronic states of the QWs overlap, the system is referred to as a superlattice (SL), which behaves as a new material electronically.

The capability of epitaxial growth to realize atomically precise hetero-interfaces has served as the basis for a number of electronic and optoelectronic device technologies including heterojunction bipolar transistors (HBTs), high electron mobility transistors (HEMTs), quantum well lasers, quantum well infrared photodetectors (QWIPs), and quantum cascade lasers (QCLs), to mention a few. In photovoltaic applications, single crystal epitaxial growth is the basis for high efficiency tandem or multijunction solar cells which hold the record for conversion efficiency as discussed earlier. They typically are designed for high performance extraterrestrial applications (spacecraft) or high performance terrestrial concentrating photovoltaic (CPV) applications. MQW systems are also of active interest for QW solar cells or several of the advanced concept devices discussed in the next section.

3.2.1.2 Nanowires

The term *nanowire* generally refers to a high aspect ratio wire-like structures in which the cross-sectional dimensions are nanometer scale, while the length may be micro to macroscale. Nanowires are generally solid, not hollow structures, the latter being referred to as *nanotubes*. Such nanowires may be oxide, metallic, or semiconducting. One of the major broad techniques used for the growth of semiconducting nanowires is *vapor-phase synthesis*, in which nanowires are grown by starting from appropriate gaseous components. In the so-called *vapor–liquid–solid* (VLS) mechanism metallic nanoparticles are used as seed sites to stimulate the self-assembled growth of nanowires. The desired semiconductor system is introduced in terms of its gaseous components and the entire assembly is heated to a temperature beyond the eutectic temperature of the metal/semiconductor system. Under these conditions, the metal forms a liquid droplet, with a typical size of a few nanometers. Once this droplet becomes supersaturated with semiconductor, it essentially nucleates the growth of the nanowire from the base of the droplet. Figure 3.6 shows examples of Si nanowires grown by VLS method using gold nanoparticles as the seeding droplets. The high-crystalline integrity of this nanowire can be clearly seen in this image, which also makes clear how the diameter of the nanowire is connected to the size of the catalyst droplet [17]. The wire shown here was

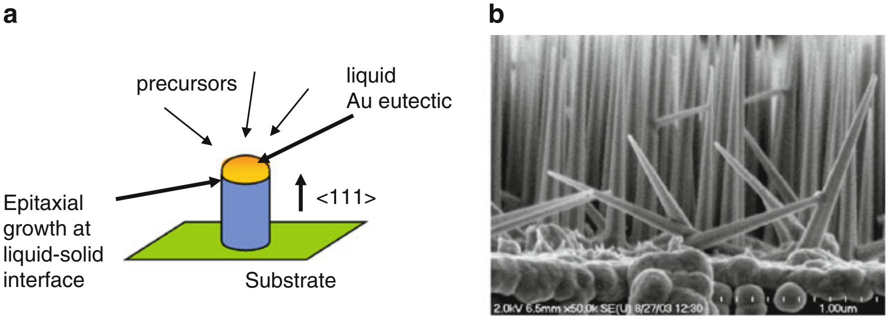


Fig. 3.6 (**a**) Schematic of the self-assembled growth of nanowires using vapor–liquid–solid (VLS) epitaxy. (**b**) Scanning electron microscopic picture of the vertical growth of Si nanowires by VLS epitaxy (from T. Picraux and J. Drucker, unpublished)

grown by using chemical-vapor deposition (CVD) to generate the semiconductor precursors, a popular approach to VLS. Other methods may also be used, however, including laser ablation and MBE. The VLS method has emerged as an extremely popular method for the fabrication of a variety of nanowires. It has also been used to realize various III–V (GaN, GaAs, GaP, InP, InAs) and II–IV (ZnS, ZnSe, CdS, CdSe) semiconductor nanowires, as well as several different wide-bandgap oxides (ZnO, MgO, SiO_2, CdO).

Samuelsson and coworkers have also had enormous success in developing nanoscale electronic devices that utilize VLS-formed, III–V semiconductor, nanowires as their active elements [18]. They have demonstrated that heterostructure nanowires of InAs and InP, as well as GaAs and InAs, can be realized that have very sharp heterointerfaces [19]. They have subsequently used this technique to implement a variety of nanoscale devices, such as resonant-tunneling diodes [20], single- [21], and multiply coupled [22, 23] quantum dots. The strong lateral confinement generated in these structures, combined with their high crystalline quality, endows them with robust quantum-transport characteristics. Quantum dots realized using these structures show very clear single-electron tunneling signatures, with evidence that the *g*-factor of the electrons can be tuned over a very wide range [24]. The ability to arbitrarily introduce serial heterointerfaces into such nanowires should offer huge potential in the future for the further development of novel nanodevices.

From the perspective of energy conversion, nanowire structures are being researched as new materials for electrochemical storage and energy conversion devices, due to the large volume ratio of these structures, which improves the catalytic performance and reaction rates, as well as providing large internal surface areas for charge storage. Within renewable energy technologies such as solar photovoltaic devices, nanowires are finding increasing use in light management, reducing the amount of light lost and allowing less material to be used for the absorption of light, hence improving efficiency and lowering material cost. Most of these efforts are in the research phase or as part of start-up ventures commercially.

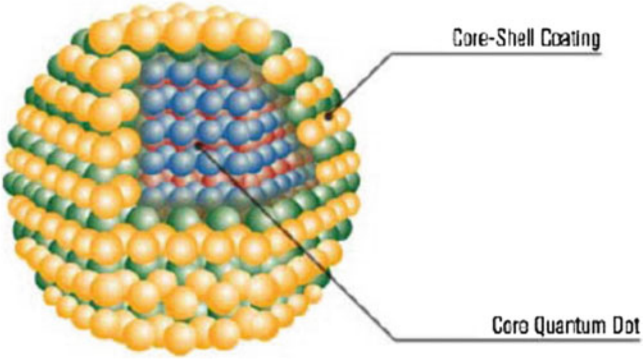


Fig. 3.7 Schematic of a core–shell nanoparticle, nanocrystal, or quantum dot structure with two different compositions

3.2.1.3 Nanoparticles and Quantum Dots

Nanoparticle is a name generally given to ultrafine size particles with dimensions on the order of 1–100 nm. If the nanoparticles are single crystal individual particles, they are often referred to as nanocrystals [25]. Alternately, agglomerates of nanoparticles are referred to as nanopowders. Nanoparticles can be metals, dielectrics, or semiconductors. They can also be grown with different compositions to form core–shell nanoparticles with unique electrical and optical properties, as illustrated in Fig. 3.7. Their electronic and optical properties are different from bulk materials as mentioned before due to quantum size effects which shift the fundamental gap to higher energy. Surface effects also play a dominant role. In particular, the dielectric properties can also be modified by surface plasma resonance effects, which change the absorption properties. The high surface-to-volume ratio affects other properties such as diffusion properties in liquid and the adhesive properties.

Nanoparticles are synthesized by a variety of techniques. One inexpensive method is through ball mill micro-machining to literally grind materials down into nanoparticles. Pyrolysis and rf plasma techniques may also be used. A popular method for synthesizing nanoparticles of high quality is through chemical solution methods, in particular sol-gel methods can realize colloidal solutions of nanoparticles which may be subsequently dried for individual nanoparticles, or the gel solutions cast for particular applications. Another method of realizing semiconductor nanoparticles is through self-assembly of InAs, or InGaAs quantum dots that on a GaAs substrate via the Stransky–Krastinov growth process [26]. In this mode of growth, a thin layer of InAs is grown on top of a GaAs substrate, but, if the layer is sufficiently thin, the strain will cause the InAs to agglomerate into small three-dimensional quantum dots.

Nanoparticles (and other nanomaterials such as nanowires and nanotubes) can be embedded in a host matrix to form a *nanocomposite*. The main differentiating factor between a nanocomposite and a normal composite material is that the large

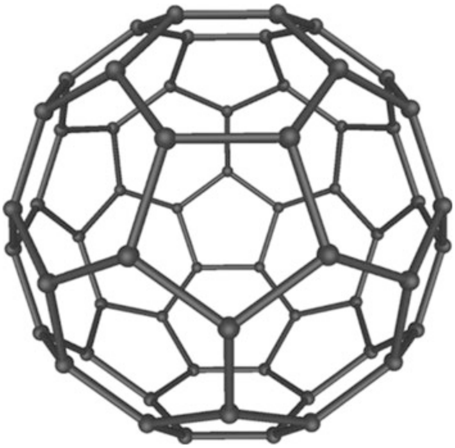


Fig. 3.8 Structure of C_{60}, Buckminsterfullerene

surface-to-volume ratio of the nanoparticle, which means that there is a large internal surface area associated with the nanoparticles compared to normal composite materials. Therefore, a much smaller amount of nanoparticle composition can have a much greater effect on the overall nanocomposite properties. Nanocomposites can be comprised of many forms, the primary ones be ceramic matrix, metal matrix, or polymer matrix nanocomposites.

3.2.1.4 Carbon-Based Materials

One of the major nanomaterials that has led to an explosion in growth in nanotechnology are those based on closed structures of graphene sheets (1 layer of graphite) composed of sp^2 bonded hexagonal rings. The term *fullerene* is used to denote any hollow closed structure such as C_{60} (Buckyballs) and CNTs. Even the properties of graphene itself have become a major focus of research due to their extraordinary electrical and thermal properties. Below we briefly review a few of the major carbon-based technologies. The first fullerene that was discovered was C_{60}, a soccer-ball-shaped object first reported by the Rice University group [27] (who were later awarded the 1996 Nobel Prize in Chemistry), the so named Buckminsterfullerene, shown in Fig. 3.8. Other such fullerenes with 72, 76, 84, etc. carbon atoms have since been synthesized. The most common method of producing C_{60} is through a carbon arc plasma between two graphite electrodes in an inert ambient. C_{60} has also found application in organic electronics and energy applications, and there it can facilitate charge transfer across interfaces. Relative to photovoltaics, they are a common component of organic photovoltaic (OPV) devices as an acceptor material or for improved charge transport.

Single-walled carbon nanotubes (SWCNTs) are a tubular form of carbon with diameters as small as 1 nm and lengths of a few nm to microns. CNTs have received considerable attention due to the ability to synthesize NTs with metallic,

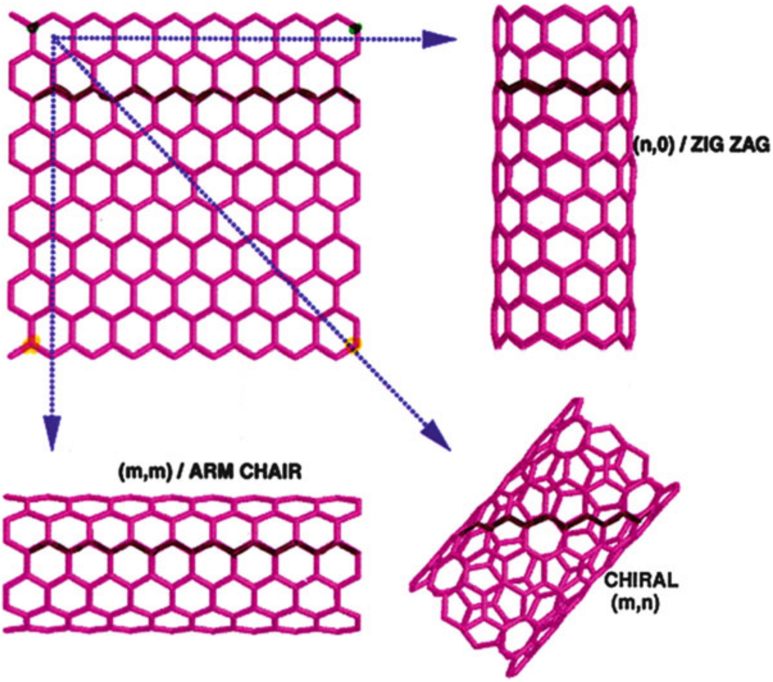


Fig. 3.9 Formation of different chirality CNTs formed by rolling a graphene sheet into a tube

semiconducting, and insulating behavior, depending on the diameter, and particularly on the chirality (i.e., how the graphite sheets forming the structure of the CNT wrap around and join themselves) [28]. Due to their remarkable electronic and mechanical properties, CNTs are currently of interest for number a of applications including interconnects; CNT-based molecular electronics; AFM-based imaging; nanomanipulation; nanotube sensors for force, pressure, and chemical; nanotube biosensors; molecular motors; nanoelectromehanical systems (NEMS); hydrogen and lithium storage; components in solar cells; and field emitters for instrumentation including flat panel displays including optically transparent films.

The structure of a SWCNT is illustrated in Fig. 3.9. It starts with a single monolayer of graphite (graphene), which has a hexagonal lattice structure characteristic of hybridized sp^2 C–C bonding. The graphene sheet may be rolled to join itself in many different ways, characterized by a pair of integers m and n, which signify the number of unit cells in the x and y direction in which the sheet is joined to itself. The larger m and n, the larger the diameter of the nanotube, with a diameter $d = a/\pi\sqrt{n^2 + nm + m^2}$, where $a = 0.246$ nm is the lattice constant. CNTs have very large aspect ratios; while the diameter may be nanometers in dimension, the length may be micron to centimeter scale in length. The ratio of m and n determines the chirality of the CNT, which determines whether it is metallic or semiconducting. *Zigzag* CNTs correspond to any combination 0, n. Another common CNT configuration is the *armchair* CNT corresponding to $n = m$. Generally speaking, armchair CNTs are metallic, with zero bandgap. Zigzag CNTs are semiconducting, unless n is a multiple of 3, with a bandgap approximately equal to 0.8/d eV, where d is the diameter of the nanotube in nanometers.

In addition to SWCNTs, multiwalled carbon nanotubes (MWCNTs) consist of multiple rolled layers (concentric tubes) of graphene. A triple walled, armchair honeycomb carbon nanotube may, for example, be made from three single-walled nanotubes. Another possible arrangement for a multiwall nanotube is by rolling a single sheet of graphene around itself, similar in shape to a snail shell. The distance between the layers is approximately 3.4 Å, which corresponds to the distance between graphene layers in bulk graphite.

In terms of their electronic transport properties, measurements on CNTs have demonstrated very high conductivities, due to the high sheet carrier density in a small dimension, the high carrier mobilities, and nearly ballistic transport [29, 30]. For this reason CNTs have been considered as viable candidates for high performance conductors and interconnects, with higher potential conductivity than copper. In terms of electronics applications, complementary n and p-channel transistors have been fabricated from CNTs, and basic logic functions demonstrated [31]. The primary difficulty faced today in a manufacturable integrated circuit (IC) technology is the directed growth and placement of CNTs with the desired chirality and diameter, suitable for large-scale production.

Typical synthesis methods include the arc discharge of graphite (the original method used to first realize CNTs), pyrolysis/thermal decomposition of a carbon source such as hydrocarbons, or laser ablation. For commercial production, the most common approach however is the catalytic growth of CNTs using chemical vapor deposition (CVD) growth, where typically metal particles are used as a catalyst. As grown, CNTs can be bundled and must be dispersed and separated to access individual nanotubes. In certain processes, like Plasma Enhanced CVD (PECVD), SWCNTs and MWCNTs grow in dense vertical arrays, with diameters and density determined by the size and distribution of the nanoparticle metal catalyst. As a result of their multifunctional properties, CNT polymer composites are expected to be used as low weight structural materials, optical devices, thermal interface materials, electric components, and electromagnetic absorption materials.

3.3 Nanotechnology and Photovoltaics

Nanostructures in solar cells have multiple approaches by which they can improve photovoltaic performance (1) new physical approaches in order to reach thermodynamic limits; (2) allow solar cells to more closely approximate their material-dependent thermodynamic limits; and (3) provide new routes for low-cost fabrication by self-assembly or design of new materials. Some of the specific advantages and disadvantages presented by nanotechnology are listed as follows:

- Range of bulk materials with proper energy gaps, catalytic properties, etc., is very limited
- Nanostructured materials allow "bandgap engineering" of electronic states and energy gaps: artificial materials
- Provide intermediate energy centers within host material
- Optical absorption can be increased, reflection and other optical losses decreased
- Improve transport and reduce scattering and energy loss
- However, higher surface-to-volume ratio means surface effects dominate: higher recombination

3.3.1 Nanostructures in Existing Solar Cell Technologies

In order to approach the ultimate efficiency limits of 70% at one sun and 86.6% at maximum concentration, new physical mechanisms are necessary. However, there is still enormous room for scope in improving existing devices. Such approaches can in general be classed in two categories: the first uses nanostructures to alter the material properties (primarily band gap) such that a more ideal structure results. For example, in tandem solar cells, the lack of a 1 eV material limits the ability to achieve tandems with a larger number of solar cells in the stack and also reduces the efficiency of existing three-junction solar cells. The use of QDs or QWs can synthesize a region of lower band gap, thus improving efficiency. In order to realize high efficiency, strain compensation for the QW/QD layer is essential. These approaches have been demonstrated in several materials, including GaAs/InAs-based materials, GaAsP-based materials, and dilute nitrides.

A more radical re-design of material properties is to use nanostructures to modify a single band gap material (e.g. Si), allowing a tandem solar cell in a single material. For example, Si can serve as the basis of a tandem solar cell by using QDs with increased band gap to form the upper junctions. Such an approach is illustrated in Fig. 3.10.

3.3.2 Multiexciton

The generation of multiple electron–hole pairs from an absorbed high energy photon provides a mechanism for increasing the efficiency of a single junction solar cell above the Shockley–Queisser limit of 33% under AM1.5 conditions, as shown conceptually in Fig. 3.11. Generation of multiple electron–hole pairs has been known in bulk materials since the 1960s in Ge and has been experimentally demonstrated in bulk silicon solar cells [32]. However, impact ionization or Auger generation processes have a low efficiency in bulk materials, and too high a threshold energy for effective utilization of the solar spectrum due to crystal

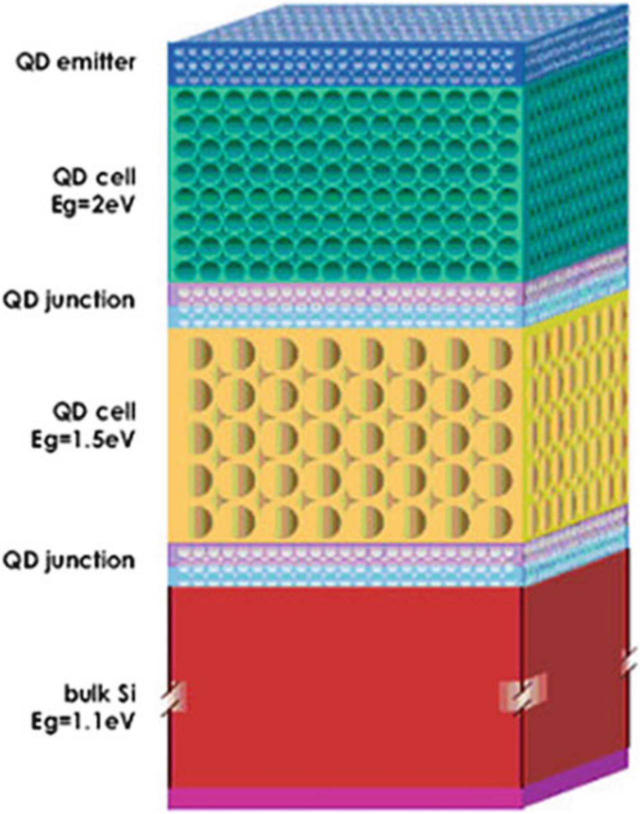


Fig. 3.10 Illustration of an all Si tandem solar cell structure based on different bandgap QD junctions (courtesy of the University of New South Wales)

momentum conservation, making the effect too small to be utilized for substantial efficiency enhancements. Nanostructured materials have been shown experimentally to increase the efficiency of carrier multiplication processes, with lower thresholds for carrier multiplication, and experimental demonstration of multiple exciton generation (MEG) in materials such as PbSe and PbS colloidal quantum dots [33, 34] with quantum efficiencies well in excess of 300%. The improved performance in nanocrystals over bulk systems is due to the relaxation of crystal momentum conservation in quantum dots, which in bulk systems together with energy conservation make the threshold for carrier multiplication roughly a factor of 1.5 higher than the bandgap. Due to quantum confinement, crystal momentum is no longer a good quantum number, and the threshold for carrier multiplication occurs at roughly multiples of the bandgap itself. Recent experimental evidence [35], as well as theoretical calculations [36], suggests indeed that the multiexcitation of several electron–hole pairs by single photons in quantum dot structures occurs at ultra-short time scales, without the necessity of impact

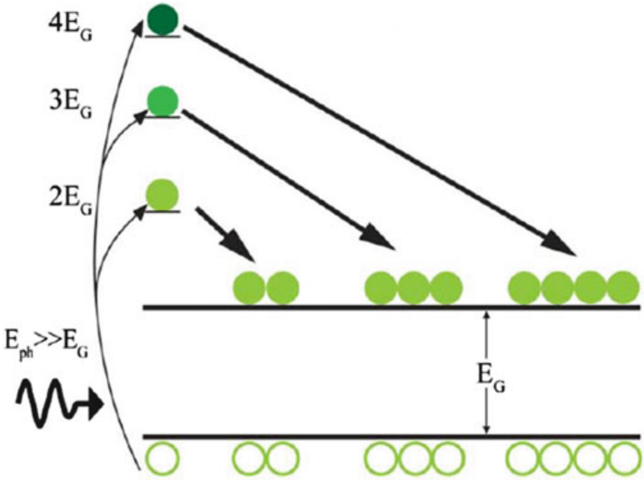


Fig. 3.11 Illustration of the multiexciton generation process for $M = 1, 2, 3$, and 4. The *lower panel* shows the calculated detailed balance efficiency as a function of bandgap with consideration of increasingly higher order multiplication factors

ionization. Overall, MEG generation has been shown in multiple materials, including PbSe, PbS, InAs [37], PbTe [38], Si [39], and CdSe [40].

While carrier multiplication has been shown in many materials, the key focus has been on lower band gap materials such as PbS and PbSe, partly because these materials show high quantum yields and partly because the theoretical optimum of a MEG solar cell is small at 0.76 eV. However, assuming more realistic conditions for carrier multiplication, the optimum band gap shifts to higher values, and Si becomes relatively ideal, as discussed below.

The increase in the efficiency of a solar cell due to MEG processes is calculated by using a quantum efficiency which is greater than unity for photon energies above the band gap. Ideally, a MEG solar cell generates two electron–hole pairs when the photon energy is between 2 and 3 times the band gap, 3 electron–hole pairs when the photon energy is between 3 and 4 times the band gap, etc. This is shown in Fig. 3.12 and in the equation below

$$Q(E) = \begin{cases} 0 & 0 < E < E_g \\ m & mE_g < E < (m+1)E_g \quad m = 1, 2, 3 \ldots \\ M & E \geq ME_g \end{cases}$$

where Q is the quantum efficiency (which is a function of energy), m is the number of electron–hole pairs generated by a photon, E_g is the threshold energy (which is ideally equal to the band gap energy), and M is the maximum number of electron–hole pairs generated.

The optimum band gap for a completely ideal MEG device is 0.76 eV [41], as shown in the detailed balance calculation of Fig. 3.13. However, the optimum low

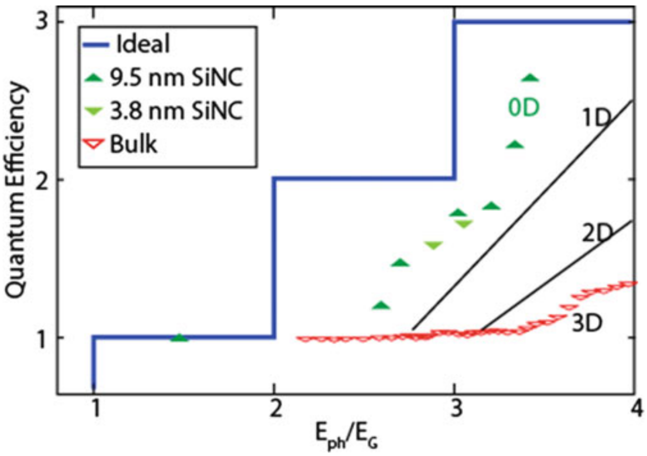


Fig. 3.12 Theoretical quantum efficiencies of a MEG solar cell, including the ideal quantum efficiency, the measured values for Si NC and bulk [40], and interpolated estimates for 1D and 2D

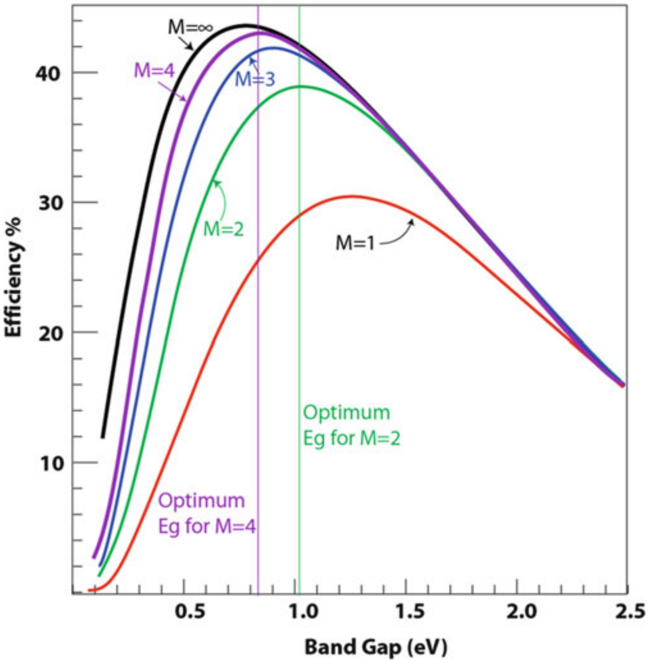


Fig. 3.13 Calculated detailed balance efficiency as a function of bandgap with consideration of increasingly higher order multiplication factors

band gap assumes the existence of multiple separate, noninteracting MEG generation processes, i.e., the band structure must be ideal for generating two excitons, as well as for three excitons, and so on, and also assumes that each of these MEG processes does not interact. For example, all photons with energy 4 times the band gap (or the threshold energy) generate 4 excitons. However, in practice, some of these photons may generate 4 excitons, while others generate 2 or 3. Finally, the ideal calculations assume that the threshold energy for all the processes is equal to m times the band gap for each one.

Assuming that in practice it is difficult to generate more than several separate exciton generation pathways, the optimum band gap is increased, and the bandgap of silicon becomes relatively ideal. This is shown in Fig. 3.13, where the optimum band gap is 1.05 eV for $M = 2$. Overall, a Si MEG solar cell with $M = 2$ achieves 90% of the ideal maximum efficiency occurring at $M = \infty$ and optimum E_G. It is interesting to note that this is a higher fraction of the thermodynamic ideal that is achieved by materials used in three junction tandems, and assuming that 78% of the ideal material-imposed efficiency limit can be achieved in a well-optimized solar cell (true for both Si and III–V cells), the Si MEG solar cell realizes an efficiency of 30% while three-junction tandems at one sun have champion efficiencies of 32%.

An additional effect which needs to be included is the increase of the effective band gap due to the quantum confinement. The effective band gap in a Si nanocrystal can be increased and tuned according to the nanocrystal size. An empirical formula governing the effective band gap is given by $E(\text{eV}) = 1.16 + 11.8/d^2$ [42]. Si nanocrystals as large as 9.2 nm display among the highest MEG effects [40], and such large nanocrystals have relatively minor effect on the effective band gap (1.3 eV as opposed to 1.16 eV). Because the maximum efficiency is not sharply peaked as a function of band gap, because the nanocrystal size may be further increased, and because the effect of stress at the nanotip peaks serves to reduce the band gap, the effect of increasing effective band gap is not considered a strong factor.

MEG processes have strong experimental verification of the effect. While initial high values measured for the MEG processes are lower under re-measurements than those initially reported, values between 130% and 300% are confirmed. These lower values substantially reduce the efficiency potential of MEG processes. However, like many of such new processes, the theoretical understanding of the effects remains incomplete, and hence, in common with other approaches, the ability to design or predict optimum structures suffers.

An approach to developing the nanostructured Si surfaces with Si nanocrystals for MEG capture is to use nanosphere lithography (NSL) due to its combination of appropriate shape, the ability to incorporate surface QDs, and the realization of narrow tips. Other approaches, for example nanoimprint lithography, have also been used to make low reflection nanostructured surfaces [43]; nanosphere masking is better integrated into a conventional solar cell process. In addition, nanostructured surfaces can be made by photolithograph-based approaches [44], electron beam lithography [45], or direct growth of nanowires [46, 47], but such approach involves either considerable cost or complexity or are not as readily incorporated into a conventional solar cell.

The approach to realizing a nanostructured Si surface using nanosphere lithography is shown in Fig. 3.14. It is based on self-assembled monolayer (SAM) of QDs

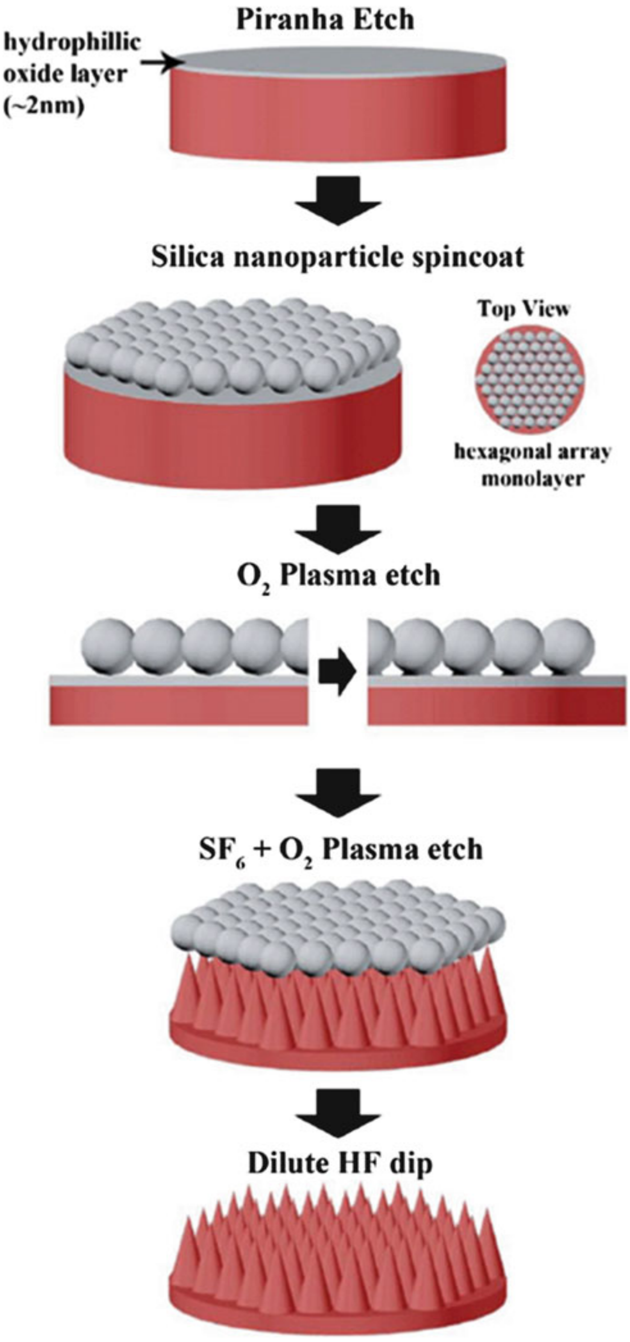

Piranha Etch
hydrophillic oxide layer (~2nm)
Silica nanoparticle spincoat
Top View
hexagonal array monolayer
O_2 Plasma etch
$SF_6 + O_2$ Plasma etch
Dilute HF dip

Fig. 3.14 Formation of nanotips of a Si surface

on the surface of a silicon solar cell. The initial demonstration of nanotips [48] uses SiO_2 nanospheres on the surfaces. These are chosen as they are available in good size uniformity, large area monolayers have been demonstrated on sapphire [49], there is good etch selectivity between SiO_2 and Si, and the SiO_2 nanospheres are stable in plasma etching. The spacing and height of the nanotips can be controlled by the size of the SiO_2 nanoparticles and the etching parameters.

The initial results demonstrate the feasibility of forming large area, defect free SiO_2 monolayers on a silicon surface, and the ability to use these spheres as an etch mask for the formation of nanotops. The surface preparation is a critical component of forming single SAMs. If the surface preparation and spincoat parameters are not optimized, then there are either areas with no nanoparticles or regions with multiple layers of nanoparticles, which interfere with the etching of the Si below. After some optimization of the surface properties and spin coat process, initial results demonstrate the ability to form large area, defect free regions with single SAM layers. Initial optimization allowed the realization of large, defect-free areas. For example, Fig. 3.15a shows a region several hundred micron on a side, with very low defects, while Fig. 3.15b shows a region with regions without SAM layers. Overall, the coverage over the entire 4 in. wafer is estimated at 60%. These results are achieved with no special modification of the surface beyond cleaning processes used in solar cell manufacturing or modification to the spin coat equipment, etc. Further, the cleaning processes did not use features such as continuous mixing, etc. which help insure the uniformity of cleans.

Initial results further show the suitability of the SAM layers as an etching mask. Figure 3.16 shows the cross-section of etched samples, with the QDs in place and with them removed following a dilute HF etch. Reactive Ion Etching (RIE) was used to etch the sample. The shape of the nanotips can be controlled by the size of the nanospheres and the etching time and properties. For example, for the given size of QDs, a shorter etching time gives a “blunter” tip due to the limited undercutting of the etching process. As the time in the RIE increases, the tips become narrower, forming 1D confinement as shown in Fig. 3.16c. The sample in Fig. 3.16c contained regions in which there were no Si QDs.

3.3.3 Intermediate Band Devices

Intermediate band solar cells, first suggested by Luque and Marti [12, 50], consist of an intermediate band between the conduction and valence band. This is shown conceptually in Fig. 3.17 for both quantum dots and quantum wells. Other realizations that have been proposed include impurity bands and bulk intermediate band materials that provide a narrow band within a larger bandgap directly from the crystal structure itself.

In order to give a thermodynamic increased efficiency, it is critical that the intermediate band be associated with its own quasi-Fermi level. If there is no separate quasi-Fermi level, then the efficiency of the overall concept collapses.

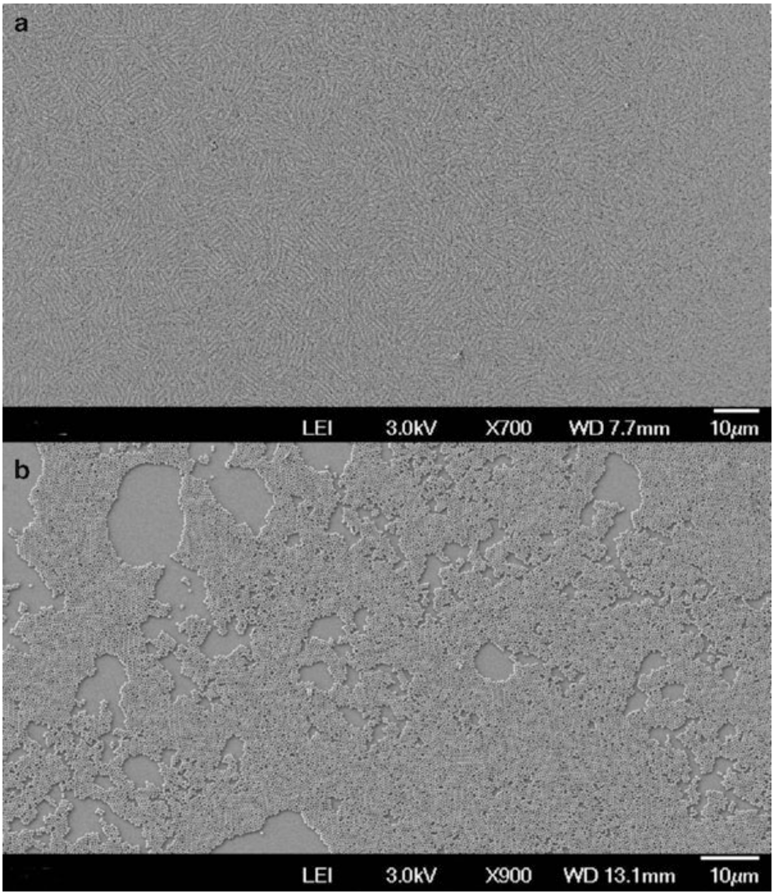


Fig. 3.15 (**a**) Large area, defect free region of single SAM coverage. (**b**) Coverage with several regions showing no SAM layers

Thus, the IB approach depends on achieving three sets of simultaneous absorption/radiative emission process (see Fig. 3.17), each having similar magnitudes, which in turn depends on realizing multiple quasi-Fermi levels. Assuming such ideal conditions, detailed balance can again be employed to calculate the maximum conversion efficiency as shown in Fig. 3.18. For any particular conduction band to intermediate band spacing, there is an optimum host bandgap giving maximum efficiency. Plotting the maximum efficiency versus the intermediate band energy, Luque and Marti were able to show that an optimum combination of host band gap (1.95 eV) and intermediate band energy (0.71 eV) leads to a maximum conversion efficiency of greater than 60% under concentration.

The three *individual* processes have a multitude of experimental evidence, with each being used in at least one commercial device [51–55]. However, the novelty and critical experimental parameter for the IB process is that the simultaneous

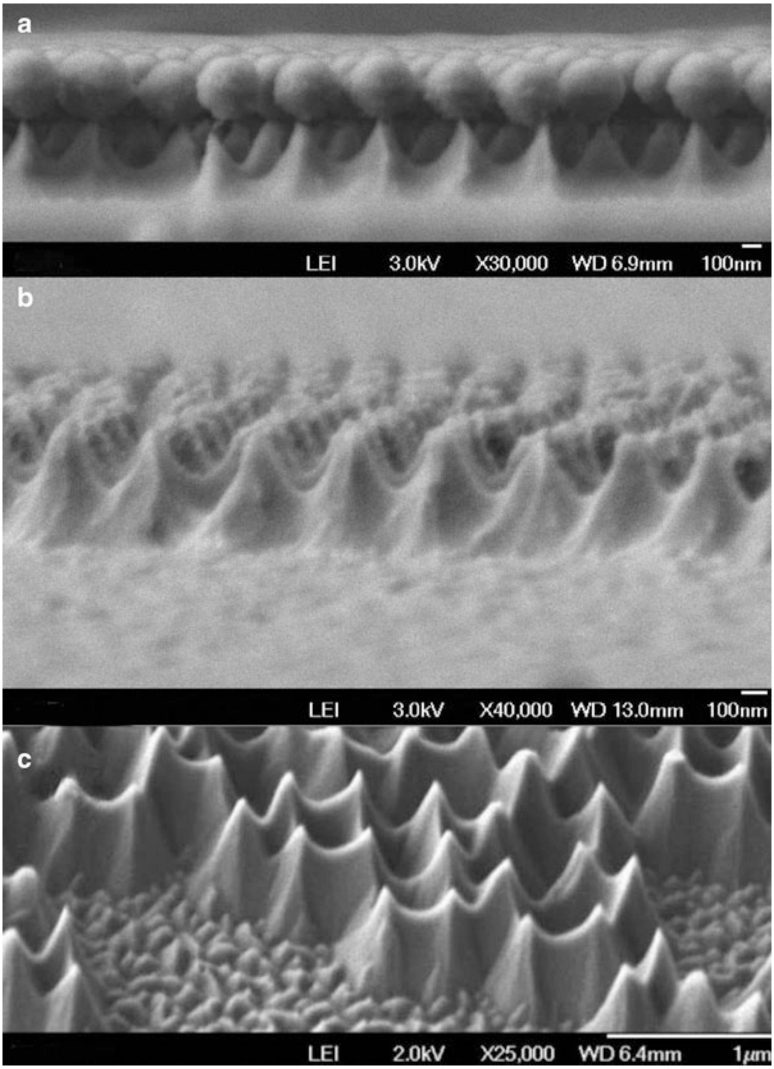

a
LEI 3.0kV X30,000 WD 6.9mm 100nm
b
LEI 3.0kV X40,000 WD 13.0mm 100nm
c
LEI 2.0kV X25,000 WD 6.4mm 1μm

Fig. 3.16 Cross section of a single SAM layer on Si (**a**) with the QDs and (**b**) with the QDs removed; (**c**) nanotips with a very sharp tip, giving 1D confinement at the tip

processes giving rise to a separate quasi-Fermi level in the intermediate band. Recently, Marti et al. have reported important results [56] showing the ability to utilize two low energy photons to collect carriers at a higher energy, a necessary prerequisite for the IB effect. Furthermore, earlier results from Nelson et al. are explained [57] by a different Fermi level inside a QW than in the barrier. Despite these important results, the demonstration of an intermediate band effect consistent with a high efficiency process remains elusive. This is due to multiple,

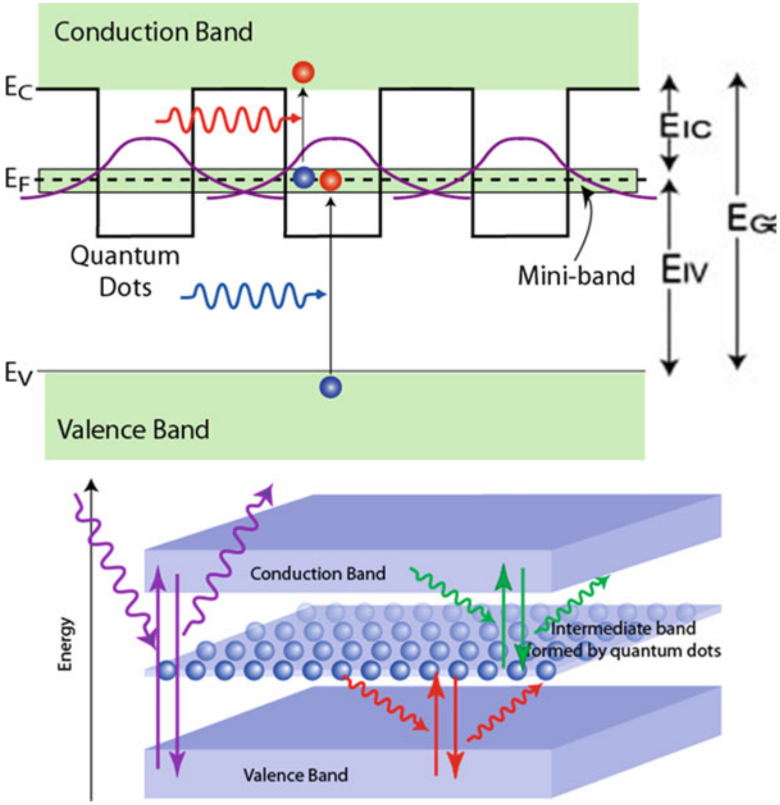


Fig. 3.17 Band diagram illustrating the realization of an intermediate band solar cell using quantum dots

interconnected reasons, relating to uncertainties about modeling and understanding the physical processes, measurement and experimental demonstration of multiple quasi-Fermi levels, identification of materials which display multiple quasi-Fermi level properties, and the design of devices based on measured properties.

In addition to the experimental demonstration of the effect, another critical issue in the intermediate band approaches is the search for optimum materials, which show appropriate band structure. While there are several bulk materials that show an intermediate band, most of the experimental work centers on using QD structures. QD structures are preferred to QW structures due to the low density of states between energy levels, which reduces scattering to lower energy levels. One limiting design rule is that the valence band offset should be small to reduce carrier recombination, increase hole collection in the contacts, and achieve a large open circuit voltage. One such material is GaAsSb/InAsP, which achieves a negligible valence band offset theoretically leads to close to ideal values for the host bandgap and intermediate band level, as shown in Fig. 3.19.

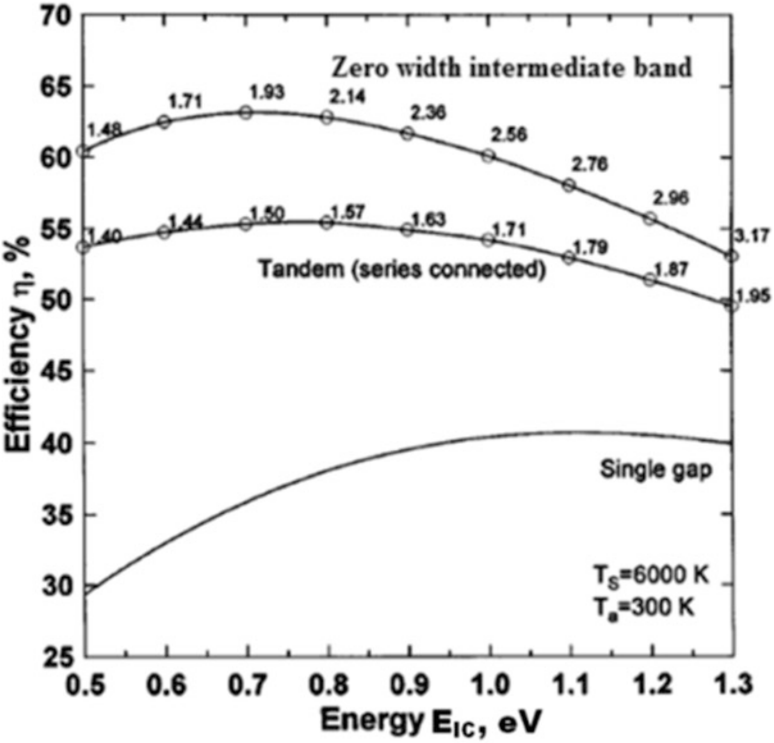


Fig. 3.18 Detailed balance efficiency as a function of the intermediate level energy (relative to the conduction band) for different host band gaps (from Luque et al. [12])

In addition to the composition and considerations of valence band offsets, the use of GaAsSb as a substrate material for the self-assembled growth of InAs quantum dots can lead to improved uniformity and reduced dot size [58]. Figure 3.20 shows the effect of different Sb mole fractions on the InAs dot density and size during epitaxial growth, showing reduction in size and increased coverage with increased Sb composition.

Another issue of importance for the operation of intermediate band solar cells in terms of the optimum quasi-Fermi level position is the appropriate doping of quantum dots, so that there is a balance of carriers excited optically from the intermediate band to the conduction band, and likewise from the valence band to the intermediate band [59]. Ideally this would correspond to half filled occupancy. Figure 3.21 shows a schematic of controllably filling the quantum dot states through a delta-doping layer of Si dopants, grown adjacent to the dots in the GaAs or GaAsSb barrier material. Time-integrated photoluminescence (PL) for samples with delta doping levels corresponding to 0, 2, 4, and 6 electrons per dot was measured as a function of both the excitation power and temperature. Typical

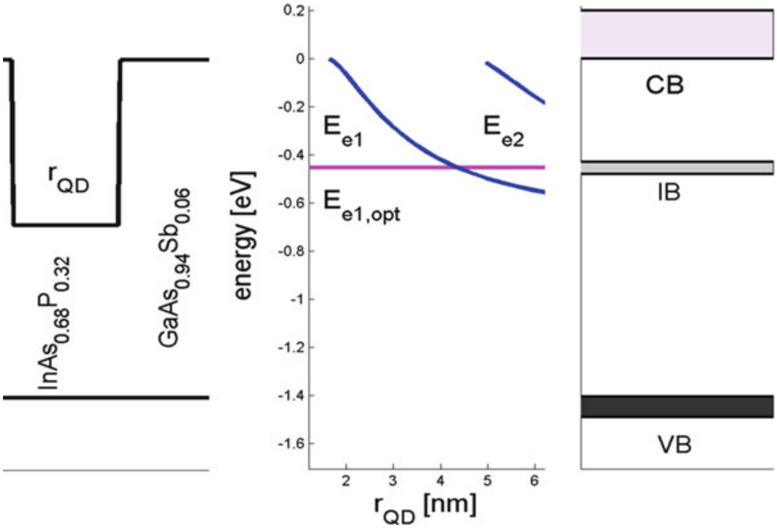


Fig. 3.19 Calculated conduction and valence band offsets for GaAsSb with InAsP quantum dots, and the corresponding energy levels for various quantum dot sizes

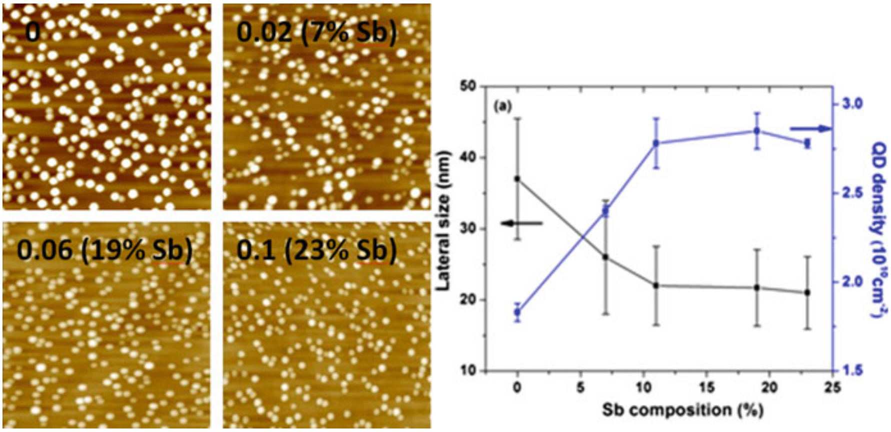


Fig. 3.20 AFM images of InAs QDs grown on the GaAs (5 ML)/GaAs$_{1-x}$Sb$_x$ (5 nm) buffer layers with various Sb compositions of 0 %, 7 %, and 23 %, respectively. *Right*: average lateral size (*left axis*), and dot density (*right axis*) of InAs QDs as a function of a Sb composition in the GaAsSb buffer layer [58]

spectra, shown in Fig. 3.22, illustrate the effect of doping on the dot PL spectra. For undoped samples, the dominant transition is from the ground state transition, whereas with increasing doping, the transition shifts to the first excited level within the quantum dot.

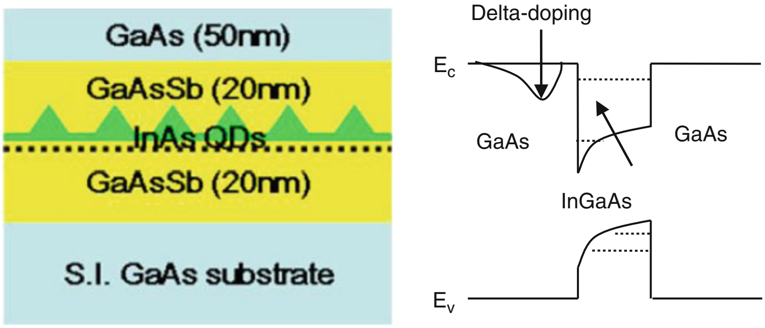


Fig. 3.21 Schematic of delta-doped InAs quantum dots in GaAsSb/GaAs for use as intermediate band states

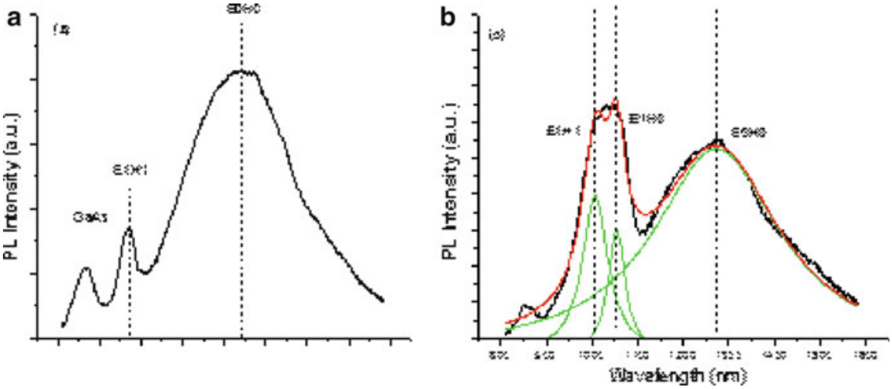


Fig. 3.22 Photoluminescence spectra from (**a**) undoped quantum dots and (**b**) Si-doped InAs QDs

3.3.4 Hot Carrier Solar Cells

Ross and Nozik proposed the concept of hot carrier solar cells [60] more than 25 years ago as a means to circumvent the limitations imposed by the Shockley–Queisser limit in terms of both the loss of excess kinetic energy and the loss of subbandgap photons. Figure 3.23 presents a schematic of the basic idea. The ideal absorber represents a material with a fundamental bandgap, $E_G \geq 0$, across which electron–hole pairs are excited by photons with energies greater than E_G. In the absorber, the relaxation of excess kinetic energy to the environment (i.e. the lattice) is suppressed, while the carriers themselves still interact strongly to establish a thermalized distribution, such that the electrons (and holes) are characterized by an effective temperature, T_H, much greater than the lattice

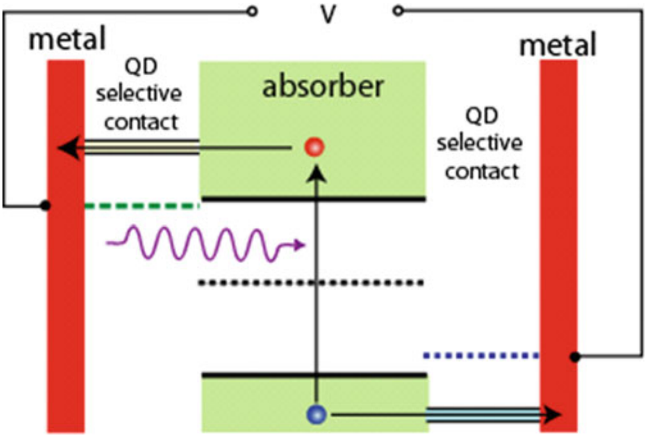


Fig. 3.23 Schematic of a hot-carrier solar cell consisting of an ideal absorber with energy selective contacts

temperature, T_L. This carrier temperature can be so large as to reverse the net chemical potential difference, μ_{ch}, between electrons and holes, and typically must be on the order of several thousand degrees for efficient operation.

Energy selective contacts are made to the absorber on the left and right, where the left contact extracts hot electrons in a narrow range of energies above the conduction band edge as shown, while the contact on the right extracts holes (injects electrons) at a specific energy range in the valence band. In this scheme, the electrons and hole are extracted from the system before they have time to relax their excess energy, hence utilizing the total energy of the photon. Under the assumption of no energy loss, the maximum efficiency occurs for vanishingly small bandgaps, hence capturing photons over the entire solar spectrum. In this limit, the theoretical detailed balance conversion efficiency approaches the maximum thermodynamic conversion efficiency of 85.4% [61]. More recently, Würfel and coworkers considered the effect of impact ionization and secondary carrier generation on the ultimate efficiency of this concept [62, 63].

There are many practical limitations to implementing this very ideal structure. One difficulty is realizing energy selective contacts. Würfel pointed out [63] that it is necessary to spatially separate the absorber material for the cold metallic contacts themselves, which may serve as an energy loss mechanism to the carriers in the absorber layer. There it was suggested that a large bandgap material such as GaN serve as a spacer or “membrane” separating the absorber from the contacts. Other proposals for energy selective contacts include using nanostructured resonant tunneling contacts from double barrier heterostructures, defects, or artificial quantum dots [64].

The main challenge in the technology is to realize an ideal absorber in which the excess kinetic energy of the photoexcited carriers is not lost to the environment.

There have been various proposals for reducing the carrier cooling rate. Due to the reduced dimensionality and therefore reduced density of final states in nanostructured systems, the energy loss rate due to phonons may be reduced, which has been observed experimentally [65]. In particular, in nanostructured systems such as quantum wells, quantum wires, or quantum dots, where intersubband spacing between levels is less than the optical phonon energy, then the optical emission rate may be suppressed due to the so-called phonon bottleneck effect, since there is no final states for the electron. However, even in such systems, the reduced phonon emission rate is still too fast for sufficient carrier heating, even under high solar concentration.

If, however, the energy is retained in the coupled electron–phonon system, then the energy would be recycled through hot phonon re-absorption. Nonequilibrium hot phonon effects during ultrafast photoexcitation have been well studied for many years. Time-resolved Raman scattering has been used, for example, to characterize the optical phonon decay after photoexcitation for a variety of III–V compound bulk and quantum well materials [66–69]. Ensemble Monte Carlo (EMC) simulation has previously been used to theoretically model ultrafast carrier relaxation and hot phonons effects in quantum well and bulk materials [70, 71], where hot phonons have been shown to significantly reduce the rate of carrier cooling compared to the bare energy loss rate.

Basically the main energy relaxation channel for electrons is through optical phonons, which lose energy through optical phonon emission in quanta of the optical phonon energy. However, due to the small group velocity of optical phonons, they do not leave the excitation volume; rather they must decay into acoustic phonons through a three phonon anharmonic scattering process, and it is the acoustic phonons which propagate energy away from the active region of the device. Hence electrons and holes may re-absorb the excess phonons, and so the excess kinetic energy of the photoexcited EHPs remains in the system until the optical phonons decay to acoustic modes. It has recently been argued by the UNSW group that nonequilibrium "hot" phonons may play a critical role in reducing carrier energy loss and maintaining energy within the absorber [72]. Typical optical phonon decay times range from 1 to 10 ps, much longer than the electron optical phonon emission rate (which is subpicosecond in scale). Engineering materials as absorbers with long phonon decays, particularly nanoengineered structures, are currently being investigated [72].

Figure 3.24 shows the simulated effect of phonon lifetime on carrier relaxation using EMC simulation, similar to earlier work on this topic [70, 73]. Here a 2 eV laser pulse exciting a 10 nm GaAs/AlAs QW is simulated, which peaks at 1 ps into the simulation, and is 200 fs wide. Optical absorption is modeled by creating electron–hole pairs corresponding to photons with a given frequency and momentum. Figure 3.24 plots the carrier temperature as a function of time for various assumed phonon lifetimes ranging from 0 (i.e. no hot phonons) to 100 ps. Additionally, one curve is included in which electron–hole scattering is suppressed.

As can be seen in the simulated results of Fig. 3.24, without hot phonons, the electrons cool rapidly and reach the lattice temperature within 5–10 ps. In contrast,

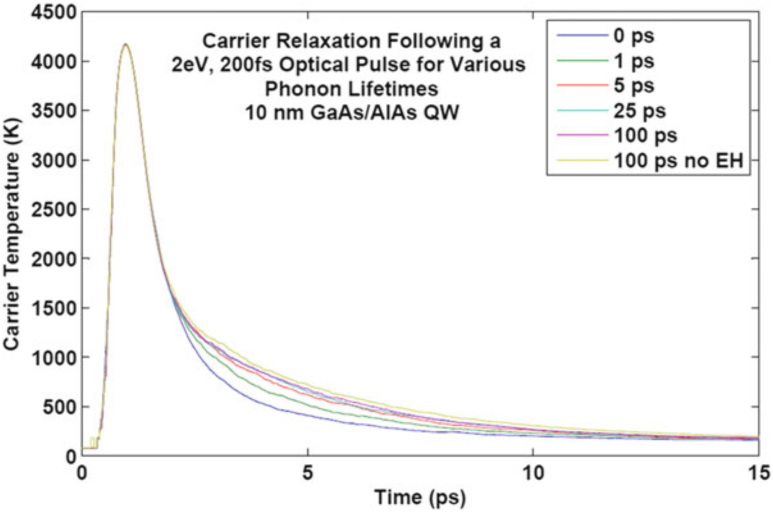


Fig. 3.24 Simulated electron temperature versus time for various assumed phonon lifetimes in a 10 nm GaAs/AlAs QW following a 2 eV, 200 fs wide optical pulse. The injected carrier density is 5 × 10^{11}/cm^2 in all cases. The lattice temperature is 5 K (S. M. Goodnick et al. [73])

with hot phonons, after the initial pulse, when a nonequilibrium distribution of hot phonons establishes itself, the decay slows and becomes nonexponential. However, for large phonon lifetimes, the decay rate does not change significantly, indicating that other channels are present for energy relaxation in the system. To check one of these, we suppress electron–hole scattering, so that energy is not taken from the electron system and transferred to the hole system, where nonpolar phonon scatter removes energy. As can be seen, the effect is not very large, and there is only a small decrease in the net relaxation rate for a phonon lifetime of 100 ps.

If the energy in the absorber is retained in the coupled electron/hole–phonon system, and only decays as the optical phonon decay, then one can look at the excess carrier energy in steady state under solar irradiation using energy balance

$$\left.\frac{\partial E}{\partial t}\right|_{\text{phonons}} + \left.\frac{\partial E}{\partial t}\right|_{\text{extr}} + \left.\frac{\partial E}{\partial t}\right|_{\text{rec}} = \left.\frac{\partial E}{\partial t}\right|_{\text{optical}} \tag{3.2}$$

The term on the right represents the average excess kinetic energy provided to the coupled electron–hole system from photoexcited carriers, whereas the first term on the left represents the energy loss rate due to optical phonons, the second term is the energy loss from the absorber due to hot carriers extracted through energy selective contacts, and the third term is the energy loss due to recombination.

The average excess energy from photons above the absorber bandgap available for carrier heating may be calculated from the appropriate solar spectrum. Figure 3.25a shows the average excess kinetic energy versus bandgap calculated from an average of the black body AM0 distribution.

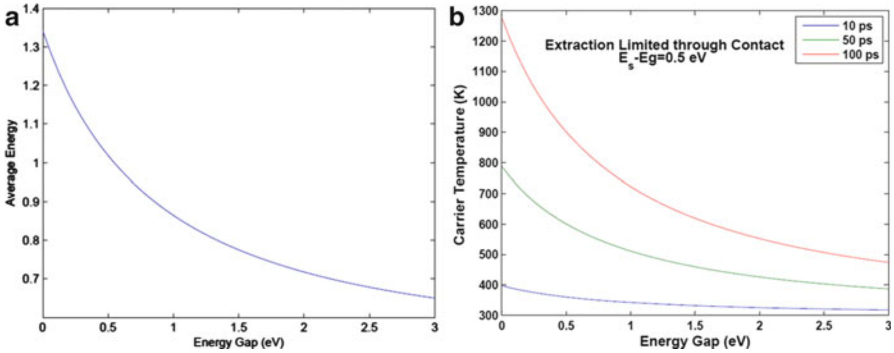


Fig. 3.25 (**a**) Average excess kinetic energy, E_{exc}, available for carrier heating as a function of bandgap based on the blackbody solar spectrum. (**b**) Calculated hot carrier temperature, T_{H}, for various phonon lifetimes assuming a 1 ns extraction time (from S. Goodnick et al. [73])

Using the result of Fig. 3.25a for E_{exc}, the calculated carrier temperature, T_{H}, as a function of bandgap for several different phonon lifetimes is plotted in Fig. 3.25b. As can be seen, for significant carrier heating to occur, long optical phonon lifetimes are required with a small gap absorber, several 10s of picoseconds, which is longer than the values measured in bulk materials, which is typically in the range from 1 to 10 ps.

The main conclusion of this analysis is that for sufficiently long energy relaxation times, hot carrier temperatures sufficient for hot carrier extraction through selective contacts are expected, leading to potentially high energy conversion efficiency. The realization of such long energy relaxation times is of course challenging and will inevitably require a combination of phonon engineering and nanostructured absorbers to suppress phonon emission.

3.3.5 Hybrid Concepts

In the previous sections, we have briefly discussed some of the advanced concept solar cell structures, and how nanotechnology can benefit the realization of these concepts. An innovation on advanced concept devices is to consider hybrids of several concepts, which allow one to exceed to potential and overcome material-specific limitations of any particular technology. Figure 3.26 shows the detailed balance calculation of the 1 sun conversion efficiency of a hot carrier solar cell versus the limits for intermediate band and multiple exciton generation solar cells, which has the potential for over 50% conversion at 1 sun, and closer to the thermodynamic limit of 86.5% at maximum concentration. Also shown is the calculated efficiency for a hybrid converter, which combines one or more concepts such as hot carrier and intermediate band [74]. As shown, the potential for such hybrid converters is greater than that of single concept approaches.

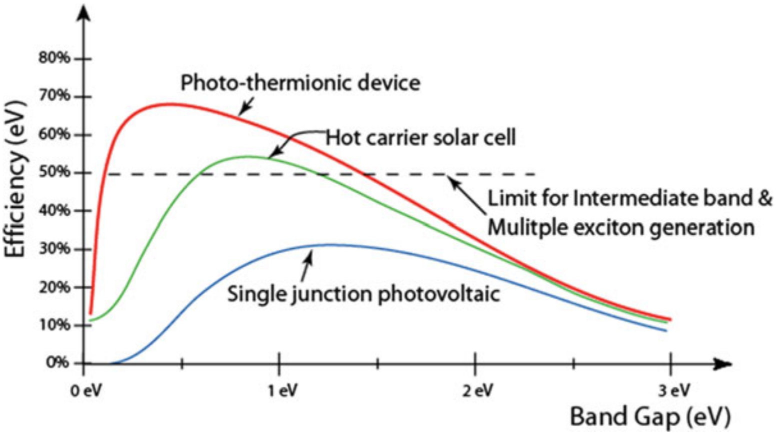


Fig. 3.26 Comparison of the detailed balance calculation of the 1 sun efficiency for a hot carrier solar cell, and a hybrid thermo-photonic device, illustrating the improvement with respect to the single gap limit

3.4 Summary

To date, many of the principles of new concept solar cells have been demonstrated as discussed earlier. However, the measured improvements in solar cell efficiency with the inclusion of nanostructures have been limited due to a variety of issues. One is the inherent problem of surfaces and interfaces in solar cells, and their passivation. Surface recombination is detrimental in terms of both reduced photocurrent affecting short circuit current and in an increase in dark current, which reduces the open circuit voltage and hence maximum power of a solar cell. Due to the high surface-to-volume ratio of nanostructured materials, they are more sensitive to surface effects than bulk materials. Therefore, effective approaches to passivate surfaces in nanostructures are necessary to incorporate them as components in the active regions of the device such as in the approaches discussed earlier. Another issue is the lack of optimization of the material structures in terms of their electronic structure and associated optical properties for the intended applications. Considerable work remains to determine the optimum nanomaterial composition and structure to realize the potential of new concept solar cells. With improvements in nanostructure growth and synthesis, it is expected that these issues will be addressed, and that benefits realized in both improved efficiency and reduction in manufacturing costs in maintaining the growth curve of photovoltaics to the TW scale.

References

1. R.E. Smalley, Future global energy prosperity: The terawatt challenge. MRS Bull. **30**, 412–417 (2005)
2. N.A. Lewis, Powering the Planet. DOE Program Review (2005)

3. W. Shockley, H.J. Queisser, Detailed balance limit of efficiency of p-n junction solar cells. J. Appl. Phys. **32**, 510–519 (1961)
4. Solar Junction tips 43.5% efficient CPV cell, preps 250 MW capacity ramp. Photovoltaics World, Photovoltaics-CPV, Issue 2, March 2011
5. R.T. Ross, A.J. Nozik, Efficiency of hot-carrier solar energy converters. J. Appl. Phys. **53**, 3813–3818 (1982)
6. K.W.J. Barnham, G. Duggan, A new approach to multi bandgap solar cells. J. Appl. Phys. **67**, 3490 (1990)
7. S. Kolodinski, J.H. Werner, T. Wittchen, H.J. Queisser, Quantum efficiencies exceeding unity due to impact ionization in silicon solar cells. Appl. Phys. Lett. **63**, 2405 (1993)
8. M.A. Green, *Third Generation Photovoltaics: Advanced Energy Conversion* (Springer, Berlin, 2003)
9. S. Kolodinski, J.H. Werner, T. Wittchen, H.J. Queisser, Quantum efficiencies exceeding unity due to impact ionization in silicon solar cells. Appl. Phys. Lett. **63**, 2405 (1993)
10. R. Schaller, V. Klimov, High efficiency carrier multiplication in pbse nanocrystals: Implications for solar energy conversion. Phys. Rev. Lett. **92**, 186601 (2004)
11. H. Cotal, C. Fetzer, J. Boisvert, G. Kinsey, R. King, P. Hebert, H. Yoon, N. Karam, III-V multijunction solar cells for concentrating photovoltaics. Energy Environ. Sci. **2**, 174–192 (2009)
12. A. Luque, A. Martí, Phys. Rev. Lett. **78**, 5014 (1997)
13. R.T. Ross, A.J. Nozik, Efficiency of hot-carrier solar energy converters. J. Appl. Phys **53**, 3813 (1982)
14. P. Würfel, Solar energy conversion with hot electrons from impact ionization. Sol. Energy Mater. Sol. Cells **46**, 43 (1997)
15. P. Würfel, A.S. Brown, T.E. Humphrey, M.A. Green, Particle conservation in the hot-carrier solar cell. Prog. Photovolt. Res. Appl. **13**, 277 (2005)
16. A.Y. Cho, J.R. Arthur, Molecular beam epitaxy. Prog. Solid State Chem. **10**, 157–191 (1975)
17. W. Lu, C.M. Lieber, Semiconductor nanowires. J. Phys. D Appl. Phys. **39**, R387 (2006)
18. L. Samuelson, Self-forming nanoscale devices. Mater. Today **6**, 22–31 (2003)
19. M.T. Björk, B.J. Ohlsson, T. Sass, A.I. Persson, C. Thelander, M.H. Magnusson, K. Deppert, L.R. Wallenberg, L. Samuelson, One-dimensional steeplechase for electrons realized. Nano Lett. **2**, 87–89 (2002)
20. M.T. Björk, B.J. Ohlsson, C. Thelander, A.I. Persson, K. Deppert, L.R. Wallenberg, L. Samuelson, Nanowire resonant tunneling diodes. Appl. Phys. Lett. **81**, 4458–4460 (2002)
21. C. Thelander, T. Martensson, M.T. Björk, B.J. Ohlsson, M.W. Larsson, L.R. Wallenberg, L. Samuelson, Single-electron transistors in heterostructure nanowires. Appl. Phys. Lett. **83**, 2052–2054 (2003)
22. A. Fuhrer, C. Fasth, L. Samuelson, Single electron pumping in InAs nanowire double quantum dots. Appl. Phys. Lett. **91** (2007). doi:10.1063/1.2767197
23. A. Fuhrer, L.E. Froberg, J.N. Pedersen, M.W. Larsson, A. Wacker, M.E. Pistol, L. Samuelson, Few electron double quantum dots in InAs/InP nanowire heterostructures. Nano Lett. **7**, 243–246 (2007)
24. M.T. Björk, A. Fuhrer, A.E. Hansen, M.W. Larsson, L.E. Jensen, L. Samuelson, Tunable effective g factor in InAs nanowire quantum dots. Phys. Rev. B **72**, 201307 (2005)
25. A.P. Alivisatos, Perspectives on the physical chemistry of semiconductor nanocrystals. J. Phys. Chem. **100**, 13226–13239 (1996)
26. D. Bimberg, M. Grundmann, N.N. Ledentsov, *Quantum Dot Heterostructures* (Wiley, Chichester, UK, 1999)
27. H.W. Kroto, J.R. Heath, S.C. O'Brien, R.F. Curl, R.E. Smalley, C60: Buckminsterfullerene. Nature **318**, 162–163 (1985)
28. M.S. Dresselhaus, G. Dresselhaus, P.C. Eklund, *Science of Fullerenes and Carbon Nanotubes* (Academic Press, New York, 1996)

29. T. Dürkop, S.A. Getty, E. Cobas, M.S. Fuhrer, Extraordinary mobility in semiconducting carbon nanotubes. Nano Lett. **4**, 35–39 (2003)
30. A. Javey, J. Guo, Q. Wang, M. Lundstrom, H. Dai, Ballistic carbon nanotube field-effect transistors. Nature **424**, 654–657 (2003)
31. P.L. McEuen, M.S. Fuhrer, P. Hongkun, Single-walled carbon nanotube electronics. IEEE Trans. Nanotechnol. **1**, 78–85 (2002)
32. S. Kolodinski, J.H. Werner, T. Wittchen, H.J. Queisser, Quantum efficiencies exceeding unity due to impact ionization in silicon solar cells. Appl. Phys. Lett. **63**(17), 2405–2407 (1993)
33. R.D. Schaller, V.I. Klimov, High efficiency carrier multiplication in PbSe nanocrystals: implications for solar energy conversion. Phys. Rev. Lett. **92**(18), 186601 (2004)
34. R.J. Ellingson, M.C. Beard, J.C. Johnson, P. Yu, O.I. Micic, A.J. Nozik, A. Shabaev, A.L. Efros, Highly efficient multiple exciton generation in colloidal PbSe and PbS quantum dots. Nano Lett. **5**(5), 865–871 (2005)
35. A.J. Nozick, Exciton multiplication and relaxation dynamics in quantum dots: Applications to ultrahigh-efficiency solar photon conversion. Inorg. Chem. **44**, 6893 (2005)
36. A. Shabaev, A.L. Efros, A.J. Nozik, Multiexciton generation by a single photon in nanocrystals. Nano Lett. **6**, 8 (2006)
37. R.D. Schaller, J.M. Pietryga, V.I. Klimov, Carrier multiplication in InAs nanocrystal quantum dots with an onset defined by the energy conservation limit. Nano Lett. **7**(11), 3469–76 (2007)
38. J.E. Murphy, M.C. Beard, A.G. Norman, S. Phillip, J.C. Johnson, S.P. Ahrenkiel, O.I. Micic, P. Yu, R.J. Ellingson, A.J. Nozik, PbTe colloidal nanocrystals: Synthesis, characterization, and multiple exciton generation. J. Am. Chem. Soc. **128**(10), 3241–3247 (2006)
39. J.H. Werner, S. Kolodinski, H.J. Queisser, Novel optimization principles and efficiency limits for semiconductor solar cells. Phys. Rev. Lett. **72**(24), 3851–4 (1994)
40. M.C. Beard, K.P. Knutsen, P. Yu, J.M. Luther, Q. Song, W.K. Metzger, R.J. Ellingson, A.J. Nozik, Multiple exciton generation in colloidal silicon nanocrystals. Nano Lett. **7**(8), 2506–2512 (2007)
41. A. de Vos, B. Desoete, On the ideal performance of solar cells with larger-than-unity quantum efficiency. Sol. Energy Mater. Sol. Cells **51**(3), 413–424 (1998)
42. T.-Y. Kim, N.-M. Park, K.-H. Kim, Y.-W. Ok, T.-Y. Seong, C.-J. Choi, G.Y. Sung, Quantum confinement effect of silicon nanocrystals in situ grown in silicon nitride films. Mater. Res. Soc. Symp. Proc. **817**, L4.3 (2004)
43. Q. Chen, G. Hubbard, P.A. Shields, C. Liu, D.W.E. Allsopp, W.N. Wang, S. Abbott, Broadband moth-eye antireflection coatings fabricated by low-cost nanoimprinting. Appl. Phys. Lett. **94**(26), 263118 (2009)
44. Y.M. Song, S.Y. Bae, J.S. Yu, Y.T. Lee, Closely packed and aspect-ratio-controlled antireflection subwavelength gratings on GaAs using a lenslike shape transfer. Opt. Lett. **34**(11), 1702–4 (2009)
45. S.A. Boden, D.M. Bagnall, Tunable reflection minima of nanostructured antireflective surfaces. Appl. Phys. Lett. **93**(13), 133108 (2008)
46. N. Wang, Y. Cai, R.Q. Zhang, Growth of nanowires. Mater. Sci. Eng. R Rep. **60**(1), 1–51 (2008)
47. M.D. Kelzenberg, D.B. Turner-Evans, B.M. Kayes, M.A. Filler, M.C. Putnam, N.S. Lewis, H. A. Atwater, Photovoltaic measurements in single-nanowire silicon solar cells. Nano Lett. **8**(2), 710–14 (2008)
48. Sean's thesis
49. T. Ogi, K. Okuyama, L.B. Modesto-Lopez, F. Iskandar, Fabrication of a large area monolayer of silica particles on a sapphire substrate by a spin coating method. Colloids Surf. A Physicochem. Eng. Asp. **297**(1), 71–78 (2007)
50. A. Luque, A. Martí, The intermediate band solar cell: Progress toward the realization of an attractive concept. Adv. Mater. **22**, 160–174 (2010)

51. Y. Yao, W.O. Charles, T. Tsai, G. Wysocki, J. Chen, C.F. Gmachl, Broadband quantum cascade laser gain medium based on a “continuum-to-bound” active region design. Appl. Phys. Lett. **96**(21), 211106 (2010)
52. P. Bhattacharya, A.D. Stiff-Roberts, S. Krishna, S. Kennerly, Quantum dot infrared detectors and sources. Int. J. High Speed Electron. Syst. **12**(4), 969–94 (2002)
53. H.F. MacMillan, H.C. Hamaker, N.R. Kaminar, M.S. Kuryla, M.L. Ristow, D.D. Liu, G.F. Virshup, J.M. Gee, 28% Efficient GaAs concentrator solar cells. IEEE Photovoltaic Specialists Conference, pp. 462–8 (1988)
54. C.G. Bailey, D.V. Forbes, R.P. Raffaelle, S.M. Hubbard, Near 1 V open circuit voltage InAs/GaAs quantum dot solar cells. Appl. Phys. Lett. **98**, 163105 (2011)
55. S. Sauvage, P. Boucaud, F.H. Julien, J.-M. Gérard, V. Thierry-Mieg, Intraband absorption in n-doped InAs/GaAs quantum dots. Appl. Phys. Lett. **71**, 2785 (1997)
56. A. Martí, E. Antolín, C.R. Stanley, C.D. Farmer, N. López, P. Díaz, E. Cánovas, P.G. Linares, A. Luque, Production of photocurrent due to intermediate-to-conduction-band transitions: A demonstration of a key operating principle of the intermediate-band solar cell. Phys. Rev. Lett. **97**, 247701 (2006)
57. J. Nelson, J. Barnes, N. Ekins-Daukes, B. Kluftinger, E. Tsui, K. Barnham, C. Tom Foxon, T. Cheng, J. Roberts, Observation of suppressed radiative recombination in single quantum well p- i- n photodiodes. J. Appl. Phys. **82**, 6240 (1997)
58. K.-Y. Ban, S.P. Bremner, G. Liu, S.N. Dahal, P.C. Dippo, A.G. Norman, C.B. Honsberg, Use of a GaAsSb buffer layer for the formation of small, uniform, and dense InAs quantum dots. Appl. Phys. Lett. **96**, 183101 (2010)
59. K.-Y. Ban, S.P. Bremner, G. Liu, S.N. Dahal, P.C. Dippo, A.G. Norman, C.B. Honsberg, Controllability of the subband occupation of InAs quantum dots on a delta-doped GaAsSb barrier. J. Appl. Phys. **109**, 014312 (2011)
60. R.T. Ross, A.J. Nozik, Efficiency of hot-carrier solar energy converters. J. Appl. Phys. **53**, 3813–3818 (1982)
61. P.T. Landsberg, G. Tonge, Thermodynamic energy conversion efficiencies. J. Appl. Phys. **5**(1), R1 (1980)
62. P. Würfel, Solar energy conversion with hot electrons from impact ionization. Sol. Energy Mater. Sol. Cells **46**, 43–52 (1997)
63. P. Würfel, A.S. Brown, T.E. Humphrey, M.A. Green, Particle conservation in the hot-carrier solar cell. Prog. Photovolt. Res. Appl. **13**, 277 (2005)
64. G. Conibeer, M.A. Green, R. Corkish, Y. Cho, E. Chob, C. Jiang, T. Fangsuwannarak, E. Pink, Y. Huang, T. Puzzer, T. Trupke, B. Richards, A. Shalav, K. Lind, Silicon nanostructures for third generation photovoltaic solar cells. Thin Solid Films **511–512**, 654 (2006)
65. W.S. Pelouch, R.J. Ellingson, P.E. Powers, C.L. Tang, D.M. Szmyd, A.J. Nozik, Comparison of hot-carrier relaxation in quantum wells and bulk GaAs at high carrier densities. Phys. Rev. B **45**, 1450–1453 (1992)
66. K.S. Tsen, K.R. Wald, T. Ruf, P.Y. Yu, H. Morkoc, Electron optical phonon interactions in ultrathin GaAs AlAs multiple quantum well structures. Phys. Rev. Lett. **67**, 2557–2560 (1991)
67. K.T. Tsen, R.P. Joshi, D.K. Ferry, A. Botcharev, B. Sverdlov, A. Salvador, H. Morkoc, Non-equlibrium electron distributions and phonon dynamics in wurtzite GaN. Appl. Phys. Lett. **68**, 2990–2992 (1996)
68. K.T. Tsen, J.G. Kiang, D.K. Ferry, H. Morkoc, Subpicosecond time-resolved Raman studies of LO phonons in GaN: Dependence on photoexcited carrier density. Appl. Phys. Lett. **89**, 112111 (2006)
69. K.T. Tsen, J.G. Kiang, D.K. Ferry, H. Lu, W.J. Schaff, H.-W. Lin, S. Gwo, Direct measurements of the lifetimes of longitudinal optical phonon modes and their dynamics in InN. Appl. Phys. Lett. **90**, 152107-1-3 (2007)
70. S.M. Goodnick, P. Lugli, Hot carrier relaxation in quasi-2D systems, in *Hot Carriers in Semiconductor Microstructures: Physics and Applications*, ed. by J. Shah (Academic, New York, 1992), pp. 191–234

71. M. Dür, S.M. Goodnick, P. Lugli, Monte Carlo simulation of intersubband relaxation in wide, uniformly doped GaAs/AlxGa1-xAs quantum wells. Phys. Rev. **B54**, 17794 (1996)
72. G. Conibeer, R. Patterson, L. Huang, J.-F. Guillemoles, D. König, S. Shrestha, M.A. Green, Modelling of hot carrier solar cell absorbers. Sol. Energy Mater. Sol. Cells **94**, 1516–1521 (2010)
73. S.M. Goodnick, C. Honsberg, Modeling carrier relaxation in hot carrier solar cells. Proc. SPIE. **8256**, 82560W (2012). doi:10.1117/12.910858
74. C.B. Honsberg, J. Lee, A. Bailey, S. Dahal, Hybrid advanced concept solar cells. Proceedings of the 37th IEEE Photovoltaics Specialists Conference, Seattle, WA, 2011

Chapter 4
Carbon Nanotube Assemblies for Transparent Conducting Electrodes

Ilia N. Ivanov, Matthew P. Garrett, and Rosario A. Gerhardt

Abstract The goal of this chapter is to introduce readers to the fundamental and practical aspects of nanotube assemblies made into transparent conducting networks and discuss some practical aspects of their characterization. Transparent conducting coatings (TCC) are an essential part of electro-optical devices, from photovoltaics and light emitting devices to electromagnetic shielding and electrochromic widows. The market for organic materials (including nanomaterials and polymers) based TCCs is expected to show a growth rate of 56.9% to reach nearly $20.3 billion in 2015, while the market for traditional inorganic transparent electronics will experience growth with rates of 6.7% to nearly $103 billion in 2015. Emerging flexible electronic applications have brought additional requirements of flexibility and low cost for TCC. However, the price of indium (the major component in indium tin oxide TCC) continues to increase. On the other hand, the price of nanomaterials has continued to decrease due to development of high volume, quality production processes. Additional benefits come from the low cost, nonvacuum deposition of nanomaterials based TCC, compared to traditional coatings requiring energy intensive vacuum deposition. Among the materials actively researched as alternative TCC are nanoparticles, nanowires, and nanotubes with high aspect ratio as well as their composites. The figure of merit (FOM) can be used to compare TCCs made from dissimilar materials and with different

I.N. Ivanov (✉)
Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
e-mail: ivanovin@ornl.gov

M.P. Garrett
Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Department of Physics, University of Tennessee, Knoxville, TN 37996, USA

R.A. Gerhardt
Georgia Institute of Technology, Atlanta, GA 30332, USA

A. Korkin and D.J. Lockwood (eds.), *Nanoscale Applications for Information and Energy Systems*, Nanostructure Science and Technology,
DOI 10.1007/978-1-4614-5016-0_4, © Springer Science+Business Media New York 2013

transmittance and conductivity values. In the first part of this manuscript, we will discuss the seven FOM parameters that have been proposed, including one specifically intended for flexible applications. The approach for how to measure TCE electrical properties, including frequency dependence, will also be discussed. We will relate the macroscale electrical characteristics of TCCs to the nanoscale parameters of conducting networks. The fundamental aspects of nanomaterial assemblies in conducting networks will also be addressed. We will review recent literature on TCCs composed of carbon nanotubes of different types in terms of the FOM.

4.1 Materials for Transparent Conductive Electrodes

A broad range of materials has been considered for possible applications as transparent conductive electrodes, from nanoparticles, graphene sheets, nanotubes (single to multiwalled), nanowires and their composites as well as composites with a polymer matrix. The decision whether to use one material or another depends on multiple factors such as the application requirements or the materials cost. More specifically, these can include work function, electron concentration/mobility, toxicity, deposition cost, and many others factors. For instance, in terms of production costs, the potential material may form a series in the order Cd < Zn <Ti < Sn < Ag < In, while in the toxicity series, the order would be different as follows Zn < Sn < In < Ag < Cd. Optimization of these cases is beyond the scope of this publication, since the goal is to give the background and references and instructional materials to facilitate research and development in the area of transparent conducting coatings (TCCs). The chapter is centered on the concept of the figure of merit (FOM) for transparent conductive coatings (TCC) for high-tech electro-optical applications. We will look at the TCE FOM evolution and show how the measurements of the FOM can be done for the nanomaterial-based TCCs. Furthermore, we explore how some details of the nanomaterial assembled structure could be derived from the measurements and also look at the effects of external doping on the FOM.

4.2 Indium: Cost, Supply–Demand Analysis

In most publications, the search for alternative transparent conductive electrodes to replace ITO is explained in terms of the increasing cost of indium. The price of indium has doubled in the last 10 years due to increasing demand, Fig. 4.1. The world production and consumption of indium was practically unchanged from the late 1980s to the early 1990s. The situation changed with the introduction of personal computers and consumer electronics to the market. In the last 20 years, there has been a continuous growth of consumption of TCCs and this has required

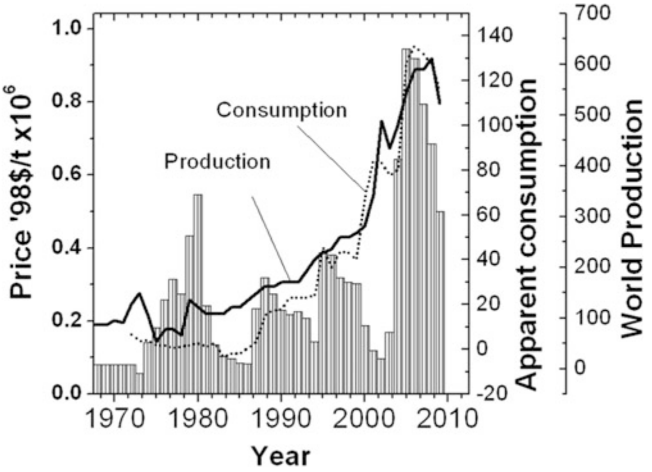


Fig. 4.1 Historical data on apparent consumption, world production, and price of indium. *Bars* show the price of indium in dollars per ton. The price is adjusted by the Consumer Price Index conversion factor, with 1998 as the base year [2]

increased production of indium. Most of indium, 70%, is consumed in coatings, and only 24% in electrical components, solder, and alloys [1]. The majority of indium is used in transparent conductive coatings. It is important that all indium used in the USA is imported from China, which supplies about 45%, Japan exporting 18%, and Canada providing 16%. Indium is a by-product of zinc mining and is imported to the USA where it is purified to electronic grade by two companies [1]. While Japan is the global leader in indium consumption, the growth of Chinese consumption has already led to cutting export quotas by 30% in the second half of 2010 to supply their domestic electronic industry demands.

As a result, 21 Chinese indium producers exported 93 tons of indium compared to the 140 tons in the first half of 2010 [1]. Indium analysts have expressed concern of possible instability of the Chinese supply to external markets. This instability of indium supply along with its increasing price has triggered significant activity in the alternative transparent conducting coatings [3–30, 50–53].

4.3 Spectral Window Considerations

The spectral selectivity of transparent conductive coatings depends on the coating functionality (application). For instance, one of the interesting opportunities to lower the power consumption of buildings is to limit solar heating through the windows, thus reducing associated air-cooling expenses by 40–70%. In the USA, this could bring millions of dollars in savings. The Department of Energy Efficiency and Renewable Energy (EERE) Building Technology Program set the goal to

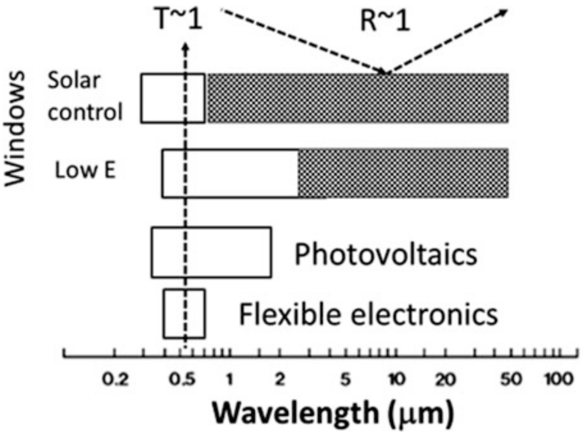


Fig. 4.2 Spectral requirements for some applications, which require TCCs. The requirement to be transparent in the visible part of the spectrum is the same for all windows and flexible displays. The reflectance and transmittance of an ideal coating, with no absorbance (100 % *T* and 100 % *R*, is referenced in the figure as *T* ~ 1 and *R* ~ 1) for visualization. More detailed requirements for different types of displays are given in the Figure of merit section

achieve 50–70% of whole building energy improvement through various approaches including new windows technology. There are two configurations of windows, which could enable the solar gain control. One type of window is the passive type with static gain control, which has continuous transmittance. The other type is the chromogenic window, which has an active or dynamic transmittance control system that allows changes in transmittance depending on the solar activity [31].

Current efforts in high-performance windows target the replacement of thick, expensive infrared reflective coatings with a thin and easy to manufacture material. Candidates include reflective transition metal hydrides or suspended particles (for more details on current R&D efforts refer to the EERE Web site: http://www1.eere.energy.gov/buildings/windows_technology.html). It should be mentioned that application of transparent conductive coatings (TCCs) in architectural windows and glass facades has a substantial market with more than 4 bl. m^2 of floating glass produced per year. The requirements for optical properties of the coating depend on each particular window technology. Figure 4.2 summarizes the transparency requirements for the various applications, which require TCCs, including architectural windows and electronics. With the solar control technology, the coatings should be transparent below 700 nm but should have high reflectivity above 700 nm. Low emissivity windows should have broader transmittance range, up to 3 μm, and high reflectivity above this wavelength. The spectral range for TCCs in a photovoltaic cell is defined by the solar spectrum (Fig. 4.3) and the optical properties of an active medium, where photocurrent is generated. Photovoltaic TCCs should have high transmittance up to 2 μm. The TCC requirements for electronics, including flexible electronics, are that it be transparent (*T* ~ 1) below 750 nm.

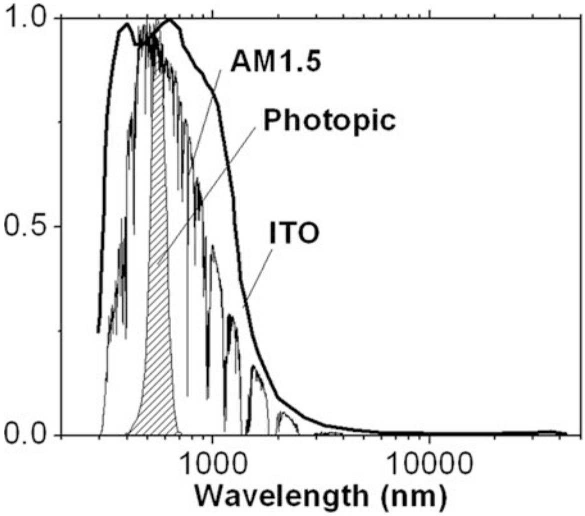


Fig. 4.3 Normalized spectral characteristics of 1.5 air mass (AM) solar spectrum relative to the averaged sensitivity of the human eye (labeled as the photopic spectral curve), and the transmittance spectrum of ITO [33]. One can see why ITO is well suited for photovoltaic applications

4.4 Why Is Transmittance of TCCs Referenced at 550 nm?

To answer this question, we have to look at the spectral characteristics of the solar spectrum and spectral response of the human eye. For photovoltaic and smart windows applications, the solar spectrum should ideally transmit all wavelengths without any losses. The shape of the solar terrestrial spectrum, which is a result of filtering through the earth's atmosphere, is shown in Fig. 4.3. An old definition of the air mass 1.5 terrestrial solar emissivity standard describes the solar spectral irradiance observed on a receiving surface at an inclined plane at 37° tilt towards the equator, facing the sun with a surface normal pointing to the sun at the elevation of 41.81° above the horizon. The old ASTM Standard (ASTM G159-98) was corrected in 2005 for deep UV (down to 280 nm, rather than 305 nm), also improving spectral resolution (a new standard uses 2002 wavelength compared to only 120 in the old version). This new standard resulted in constant intervals, better defined extraterrestrial spectrum and other factors while maintaining integrated irradiances standards for hemispherical and direct normal on 37° tilted surface at 1,000.4 and 900.1 W m^{-2}, respectively [32].

Electro-optical applications of TCCs in displays have different spectral requirements, which also depend on the spectral response of the human eye. With two basic retinal receptors: 2-μm diameter cones (concentrated in the center of the retina) and rods, our eye uses the first to sense in bright-light conditions and the second for low-intensity light conditions. The eye spectral response is directly related and influenced by the illuminescence level (light intensity) to which it is exposed. For the condition of illuminance levels (10^{-2}–10^{8} Cd m^{-2}) correlated to indoor-sunny day illuminance condition, the human vision depends mostly on the cones spectral response, which is described by the photopic vision mode for bright light (>1 Cd m^{-2}) and mesopic vision mode for dimmer light conditions. The best color and visual acuity corresponds to illumination levels of 10^{4} Cd m^{-2} for the photopic vision mode and cones sensing. The maximum of the photopic spectral responsive curve occurs at 555 nm, as shown in Fig. 4.3, for high light intensity levels. The FOM for TCCs is therefore based on transmittance/absorbance of the material at 550 nm. One should keep in mind that the spectral sensitivity shifts from ~400–730 nm to ~370–650 nm when transitioned to a low level of luminance, with corresponding transitioning of the maximum spectral response to 507 nm [34].

4.5 Development of the Figure of Merit for Transparent Conductive Electrodes

The development of the FOM is an important part of any technology, including TCC. However, in most cases, the FOM is very specific to an application, and often times it is expressed as a range of suitable properties rather than as a single number. Application-specific requirements may include environmental, mechanical stability (bending and stretching). The subject of carbon nanotube based TCC reduces FOM to the electro-optical properties of the nanomaterials. However, given the significant interest in the field of flexible electronics, we will bring an example of integration of mechanical stability into the FOM of TCCs.

Figure 4.4 shows resistance–transmittance characteristics for a series of common transparent conductive coatings, which are currently used along with ITO.

In the 1970s, Haacke developed the definition of the FOM for TCCs [35, 36].

The sheet resistance, R_s, can be expressed in terms of electrical conductivity σ (Ω^{-1} cm^{-1}) and d is the coating thickness (cm),

$$R_s = \frac{1}{\sigma d}. \tag{4.1}$$

The optical transmittance, T, of a thin film is given by the ratio of the radiation passed through the thin film (I) to the radiation entering the film (I_0) or it can also be given in terms of the absorption coefficient α (cm^{-1}) and the film thickness (d):

$$T = \frac{I}{I_0} = \exp(-\alpha d). \tag{4.2}$$

The definition of the first FOM, is a ratio of the sample transmittance to the sheet resistance, $\mathrm{FOM}_1 = \frac{T}{R_s}$, which can be rewritten in terms of the electrical conductivity, absorbance, and TCC thickness by substituting the T and R_s values from Eqs. (4.1) and (4.2):

$$\mathrm{FOM}_1 = \sigma d \exp(-\alpha d). \tag{4.3}$$

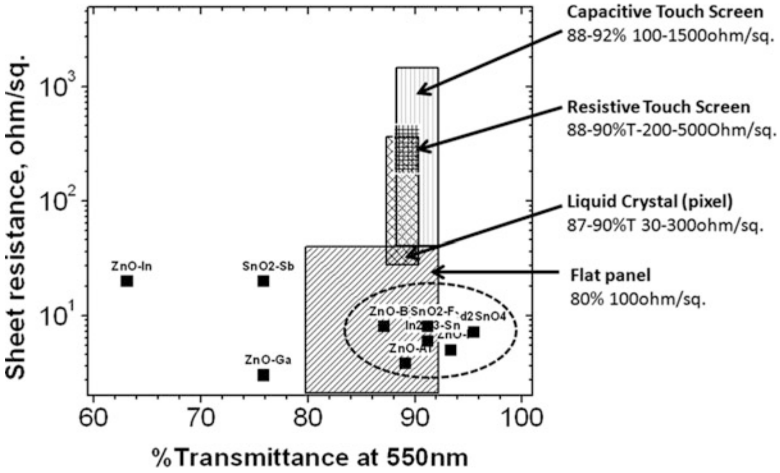


Fig. 4.4 Resistance–transmittance characteristics of some transparent conductive coatings, data adapted from [37]. The materials are produced by spray pyrolysis (SnO_2:F), sputtering (In_2O_3:Sn), chemical vapor deposition (ZnO:Ga), or by pulsed laser deposition (Cd_2SnO_4). A group of materials approaching ITO 90 % transmittance and 10 Ω/sq. sheet resistance is *circled*. The requirements for particular display applications of transparent coatings are shown in *blocks*. The toughest requirements are for flat liquid crystal display (LCD) applications, where the transmittance should exceed 87 % and the sheet resistance should be about 30 Ω/sq

Equation (4.3) expresses the FOM for a coating with conductivity σ, absorbance α, and thickness d. The maximum value of FOM according to the definition can be found taking the first derivative of (4.3), equating it to zero, and solving it for minimum thickness $d_{\min} = 1/\alpha$:

$$\frac{\partial \mathrm{FOM}_1}{\partial d} = \frac{\sigma \exp t(\propto d) - \frac{\sigma}{\alpha} \exp t(\alpha d)}{\exp t(2\alpha d)} = 0. \tag{4.4}$$

And the transmittance at maximum FOM is then found to be, $T = 1/e = 0.37$. This exercise [Eq. (4.4)] demonstrates that the original approach is not adequate to define the FOM, because the best coating would transmit less than 40% of the incident light.

To circumvent this problem, Haacke proposed another definition of FOM for transparent coatings:

$$\mathrm{FOM}_2 = \frac{T^x}{R_s}, \quad \text{where } x > 1. \tag{4.5}$$

By solving the equation $\frac{\partial \mathrm{FOM}_1}{\partial d} = 0$ for the minimum thickness required to achieve the maximum value of FOM_2, $d_{\min} = 1/x\alpha$. Through the modeling of different values of x, Haacke selected $x = 10$ as the value leading to $T = 90\%$ [35]. One should notice that $x = 20$ or 100 would lead to transparency of thin coatings of $T = 95\%$ and 99%, respectively, which are not practical. Thus, the definition of FOM_2 was written for $T = 90\%$ as:

$$\mathrm{FOM}_2 = \frac{T^{10}}{R_s} = \sigma d \exp(-10\,\alpha\, d), \tag{4.6}$$

where the corresponding minimum thickness of the coating, $d_{\min} = \frac{1}{10\alpha}$.

The right part of Eq. (4.6) was developed for TCCs with no reflection loss. If reflectivity (R) is nonzero, Eq. (4.6) changes to:

$$\mathrm{FOM}_2 = \sigma d \left\{ \left(1 - R^2\right) \left[exp(\alpha d)^{-1} \right]^{10} \right\}. \tag{4.7}$$

Equation (4.7) is a more general expression for FOM_2 because it is valid for both $R > 0$ and $R = 0$. This definition of FOM is still active. Different TCCs can be compared by their conductivity/resistance on a condition that the coatings have same transmittance (90%) at 550 nm.

Haacke noticed that between metal and semiconductor materials, the highest FOM was obtained for semiconducting coatings [35]. This was explained through the following analysis.

The ratio of electrical conductivity to absorbance in terms of the velocity of light (c), the index of refraction (n), the light frequency (v), and the relaxation time of charge carriers (τ) is given by Eq. (4.8) left part. Further, by replacing carrier lifetime $\tau = l/v$ is by the free mean path (l), it becomes apparent that the highest ratio will be obtained for the material with the largest carrier mean free path [right part of Eq. (4.8)]:

$$\mathrm{FOM}_3 = \frac{\sigma}{\alpha} = \pi c n v^2 \tau^2 = \pi c n l^2. \tag{4.8}$$

The carrier mean free path for semiconductors is on the order of 10^{-6} cm or higher compared to that of metals. Moreover, metals absorb strongly in the visible part of the spectrum, and at a thickness of a film less than 1 μm, the diffuse scattering of the charge carriers at the surface of the TCC is high, thus, significantly reducing the mean free path. Rewriting Eq. (4.8) for semiconductors, gives:

$$\frac{\sigma}{\alpha} = \frac{\pi c n v^2}{e} \mu^2 m_{\mathrm{eff}}^2 \sim A(m_{\mathrm{eff}})^{-x} (m_{\mathrm{eff}})^2 = \mathrm{const}\left(m_{\mathrm{eff}}{}^{0.64(75)} \right), \tag{4.9}$$

where charge mobility μ is replaced by $\mu = e\tau/m_{\mathrm{eff}} \sim (m_{\mathrm{eff}})^{-x}$ where e and m_{eff} are the electron charge and its effective mass and $x = 1.35$ for many materials. Equation (4.9) gives a recipe for semiconducting electrodes with high FOM i.e., those materials with high mobility or low effective mass.

Gordon continued the approach of using σ/α as the FOM and incorporated the theory of electrons in metals, which enables estimation of the theoretical upper limit of the FOM [37].

$$\mathrm{FOM}_4 = \frac{\sigma}{\alpha} = -\{R_\mathrm{s}\ln(T+R)\}^{-1} = 4\pi^2\varepsilon_0 c^3 n\lambda^{-2}\mathrm{e}^{-2}(m_\mathrm{eff}\mu)^2, \tag{4.10}$$

where ε_0 is the permittivity of free space and λ is the visible wavelength of light. The highest value of FOM_4 is expected for materials that demonstrate a high value of the product of the effective electron mass and charge mobility.

The electron-scattering processes significantly reduce the electron mobility. These processes include scattering electrons by phonons (the dominant process in single crystals) and by grain boundaries (dominant in polycrystals). In the case of doped semiconducting TCCs, scattering by ionized dopants may dominate, limiting the charge carrier mobility.

George Grüner's group suggested [3] that the correlation between transmission (in the visible spectrum) and the sheet resistance R_s should follow a metallic skin-effect model for electromagnetic waves in thin metal films by replacing the transmittance with an expression for thin metal films, which was proposed earlier by Tinkham [38]. Assuming that the TCC film thickness is smaller than the wavelength of light and neglecting the imaginary part of the conductivity, the transmittance in the visible part of the spectrum can be written as:

$$T = \frac{1}{\left(1+\frac{2\pi}{c}\sigma_\mathrm{opt}d\right)^2 + \left(\frac{2\pi}{c}\sigma_\mathrm{dc}d\right)^2} = \frac{1}{\left(1+\frac{2\pi}{cR_\mathrm{s}}\frac{\sigma_\mathrm{opt}}{\sigma_\mathrm{dc}}\right)^2}, \tag{4.11}$$

where $\sigma_\mathrm{dc}, \sigma_\mathrm{opt}$ are the DC and the optical conductivity of a film with thickness d. The ratio of DC to optical conductivities was assumed to be constant for different film thickness and equal to one for transmittance measured at 550 nm [3]. For a broad frequency range (DC to optical frequency), the ratio of $\frac{\sigma_\mathrm{opt}}{\sigma_\mathrm{dc}}$ was found to be around 3 [4].

The ratio of DC to optical conductivity, another expression of FOM for TCCs, can be written as:

$$\mathrm{FOM}_5 = \frac{\sigma_\mathrm{dc}}{\sigma_\mathrm{opt}} = \frac{2\pi}{cR_\mathrm{s}\left(\frac{\sqrt{T}}{1-\sqrt{T}}\right)}. \tag{4.12}$$

Thus, the sheet resistance of a thin film is expected to depend on its transparency as:

$$R_\mathrm{s} = \frac{\frac{c}{2\pi\left(\frac{1-\sqrt{T}}{\sqrt{T}}\right)\sigma_\mathrm{opt}}}{\sigma_\mathrm{dc}}, \tag{4.13}$$

and a plot of the sheet resistance as a function of $\left(\frac{1-\sqrt{T}}{\sqrt{T}}\right)$ is expected to be linear with a slope of $\frac{c}{2\pi}\frac{\sigma_\mathrm{opt}}{\sigma_\mathrm{dc}}$.

All FOMs described so far are applicable only to the low frequency region, where conductivity is not complex or frequency dependent. At higher frequencies, simple relationships between conductivity and transmittance are no longer valid and have to be defined by either using the Airy formula or the Kramers–Kronig transformation. Using this approach, a new FOM has been proposed by Kamaras [5]. This FOM relies on a wavelength-dependent general FOM for thick and thin nanotube TCCs. The major argument is that this model takes into account that the optical properties in the visible spectral range depend on the concentration of bound charge carriers, while DC resistivity depends on the free carriers. Applying Kramer–Kronig transformations for free standing laser ablation nanotubes, they demonstrated that due to metallic and semiconducting nanotubes contributing to absorbance at 550 nm, the optical conductivity derived from Grüner's approach [3] overestimates it by a factor of 2. They also showed that the optical conductivity is wavelength (frequency) dependent. Therefore, the ratio of $\frac{\sigma_{\text{opt}}}{\sigma_{\text{dc}}}$ and Eq. (4.11) should be frequency dependent:

$$T(w) = \frac{1}{\left(1 + \frac{2\pi}{cR_{\text{s}}} \frac{\sigma(w)_{\text{opt}}}{\sigma_{\text{dc}}}\right)^2}, \tag{4.14}$$

where w is the wavenumber in cm^{-1}.

The authors found that the optical properties of nanotubes could be fitted by the Drude–Lorentz model, where the Drude part models the contribution of the metallic nanotubes (free charge carriers). Based on the proposed model and deconvolution of the transmittance spectra, the transmittance in Eq. (4.2) describes the far-infrared free charge carrier for $w > 2{,}000$ cm^{-1} (below 5,000 nm), and the optical conductivity of these charge carriers is zero.

The following assumptions are introduced: the mass of electrons is the same as the mass of charge carriers, the DC conductivity can be obtained by the zero limit of the optical conductivity, the frequency w is less than the relaxation rate of free charge carriers, and that in the frequency range of the Lorentzian contribution, the reflectance of the film is negligible, and the absorbance obeys Beer's law. Although no assumption is made about interrelation of the number of bound and free charge carriers, one can assume that they should be proportional to the number of metallic N_1 and semiconducting N_2 nanotubes, and for laser-grown carbon nanotubes, the ratio of N_1/N_2 is expected to be around 1/3. The optical conductivity, described by the Drude–Lorentz model, can be written as:

$$\sigma(w)_{\text{opt}} = \frac{N_1 e^2}{mV} \frac{\gamma_1}{(w^2 + \gamma_1^2)} + \frac{N_2 e^2}{mV} \frac{\gamma_2^2}{(w_0^2 - w^2)^2 + \gamma_2^2 w^2}, \tag{4.15}$$

where V is the volume of the system, N_1, m, and γ_1 are the number of free charge carriers, the mass of the charge carriers, and the width of the free carrier conductivity (the relaxation rate), respectively. The same quantities labeled with subscript

2 correspond to bound charge carriers. In the zero limit, and for frequencies $w < \gamma_1$, Eq. (4.15) expresses the sheet conductance of nanotubes, σ_s:

$$\sigma_s = \frac{N_1 e^2}{mV\gamma_1} d. \tag{4.16}$$

Beer's law for Lorentzian contribution can be expressed in terms of the frequency-dependent extinction coefficient $\alpha(w)$, as follows:

$$-\log(T(w)) = \frac{a(w)N_2}{V} d. \tag{4.17}$$

Combining this equation and Eq. (4.16), we see that the value of σ_s depends linearly on $-\log(T(w))$ with a slope, which expresses the ratio between the number of free charge carriers responsible for conductance to the number of bound charge carriers responsible for high frequency absorbance. This slope is used by Kamaras as the "ideal" FOM expression.

$$\mathrm{FOM}_6^{\mathrm{ideal}}\ \mathrm{FOM}_6 = \frac{\sigma_s}{-\log(T(w))} = \frac{N_1\ e^2}{N_2 \in (w) m\,\gamma_1} \tag{4.18}$$

Although the expression of FOM for the nonideal case, where DC conductivity is determined by the contacts (this does not meet the assumption of the zero frequency for optical conductance), was also derived, we omit it here because information about the charge carrier concentration cannot be derived directly from the experimental data and the description is beyond the scope of this manuscript.

The FOM_6 was derived to reflect the fact that experimental data do not go through the origin, but they intercept the absorbance axes at zero conductivity. This value of transmittance, $T_p(w)$ is related to the percolation threshold condition and is frequency dependent and has also been observed by Hu and Gruner [3]. Applying this correction to the expression for FOM_6 gives

$$\mathrm{FOM}_6 = \frac{\sigma_s}{-\log T(w)/T_p(w)}. \tag{4.19}$$

Based on this definition of FOM_6, unsorted CoMoCat[1] nanotubes [two grades commercial and enriched with (6, 5) and (7, 5) tubes], ark, and HiPCO[2] single-

[1] CoMoCat SWNTs are grown by CO disproportionation on Co–Mo catalyst at temperatures 700–950 °C at pressure that ranges from 1 to 10 atm. The process was developed by Prof. Daniel Resasco group at the University of Oklahoma and was successfully scaled up by South West Nanotechnologies (SWeNT) for production of single- and multiwall carbon nanotubes (CNT), http://swentnano.com/tech/what_is_comocat.php.

[2] HiPCO-SWNTs are grown by high-pressure carbon monoxide (CO) processing at high temperatures. The growth of HiPCO SWNTs was developed by Dr. Richard Smalley's group, which also demonstrated scaled up their production of nanotubes. The process uses iron carbonyl as catalyst, which reacts with carbon monoxide gas in a chamber at high pressure and high temperature.

walled carbon tubes TCCs were compared (Fig. 6 in [5]). The comparison was conducted for 550 and 1,600 nm wavelengths (see Table 1 in [5]). The nanotubes formed the following series HiPCO, CoMoCat (commercial), arc, CoMoCat [enriched with (6, 5) and (7, 5) tubes], with HiPCO tubes showing the highest FOM_6. The numerical values of the best and worst FOM_6 for 550 nm are 0.0068 and 0.0013, respectively.

Analysis of FOM for HiPco, laser ablation-grown, and arc discharge SWNTs using optical conductivity measured across the UV-NIR spectrum gave a very different result. The sorted and laser-produced nanotubes showed the highest FOM_5: sorted and laser ablation-grown > arc discharge > HiPco SWNTs [17]. The observed difference may be related to a variety of factors, including different lengths of nanotubes and bundle-to-bundle resistance.

Flexible electronics require not only transparency and low resistivity but also flexibility of the electrode material. All the FOMs described above do not consider flexibility, focusing only on the transparency and the conductivity aspect. Kotov's group proposed an idea for further modification of the FOM to reflect the mechanical stability of flexible electrodes [39]. This particular FOM may be beneficial for touch screen displays, wearable sensors, and many others.

$$\mathrm{FOM}_7 = \frac{\sigma \varepsilon_\mathrm{c}}{\alpha},\ \mathrm{ohm}^{-1}, \tag{4.20}$$

where ε_c is the critical strain before critical failure, which can cause a device to malfunction. The strain of a component layer in a film can be estimated as a ratio of the distance from the neutral axis layer and the radius of the bending curvature. The value of ε_c depends on the thickness of the substrate, the buffer layer adhesion of a coating, and many other factors. To measure the ε_c, one could find the point of inflection on a strain–resistance curve or measure the onset of crack formation on the film's surface. It is interesting to note that the value of ε_c (for LbL SWNT films) was 99 and 120% before and after super acid doping, while the critical strain for ITO coating on PET was 1% for tension and 1.7% for compression. Thus, the SWNT film demonstrates 100 times improvement in bending properties as compared to ITO, if the coatings are compared using the critical strain values. However, the difference is only a factor of 2 if the FOM_6 for these coatings are compared. The $\mathrm{FOM}_7 = 0.15\ \Omega^{-1}$ for the layer-by-layer assembled SWNT film, while the FOM_7 (ITO) $= 0.07\ \Omega^{-1}$. The value of the critical strain before critical failure in (4.20) could be replaced by other mechanical property, which is more critical for the particular TCC application (for instance critical tensile strength, toughness, Young's modulus can be used). One should probably consider a series of experiments, where both ITO and competing nanomaterial coatings are deposited on exactly the same substrate, using the same fabrication method and tested under similar conditions. Table 4.1 summarizes the definitions of FOMs described above and what conditions they apply under.

Figure 4.5 summarizes most of the known reports on nanotube-based TCCs, where a series of coatings were tested to demonstrate a percolation behavior for

Table 4.1 The expressions of the figure of merit for transparent conductive coatings

FOM	Expression	Notes
1	$\sigma d\exp(-\propto\ d)$	Maximum FOM_1 at $T = 37\%$
2	$\frac{T^{10}}{R_s} = \sigma d\exp(-10\alpha d)$	No reflectance
	$\sigma d\{(1-R^2)\Big[\exp(\propto\ d) - R^2\exp(-\propto\ d)^{-1}\Big]^{10}\Big)$	General expression, includes reflectance term, R
3	$\frac{\sigma}{\alpha} = \pi cnv^2\tau^2 = \pi cnl^2$	l is the carrier lifetime
4	$\frac{\sigma}{\alpha} = -(R_s\ln(T+R))^{-1} = 4\pi^2\varepsilon_0 c^3 n\lambda^{-2}\mathrm{e}^{-2}(m_{\mathrm{eff}}\mu)^2$	n is the index of refraction
5	$\frac{\sigma_{\mathrm{dc}}}{\sigma_{\mathrm{opt}}} = \frac{2\pi}{cR_s\left(\frac{\sqrt{T}}{1-\sqrt{T}}\right)}$	For TCC with the thickness less than the wavelength of light
6	$\frac{\sigma_s}{-\frac{\log T(w)}{T_p(w)}}$	$T_p(w)$ is related to the percolation threshold condition and is frequency dependent
7	$\frac{\sigma\varepsilon_c}{\alpha}$	For flexible TCC. ε_c-critical strain before critical failure

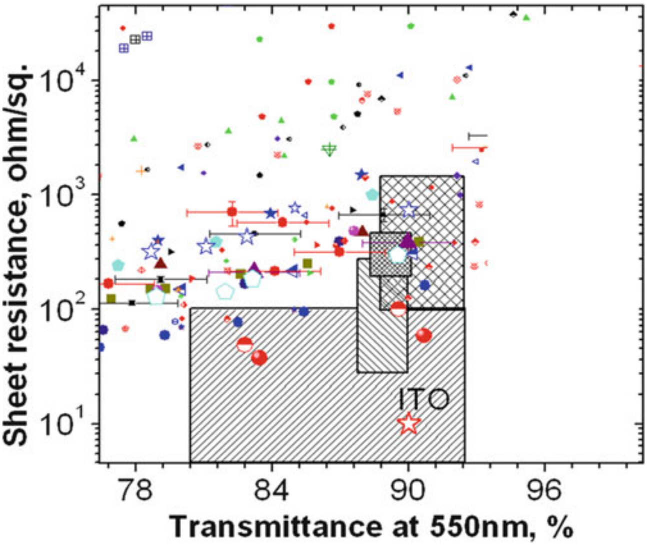


Fig. 4.5 A summary of the electro-optical properties of carbon nanotube-based (SWNT, DWNT, and MWNTs) transparent conductive coatings reported in the literature. The *grayed areas* show possible applications including resistive and capacitive touch screen, liquid crystal, and flat panel displays. More details on the requirements for each particular display application are shown in Fig. 4.4. Carbon nanotube coatings already meet the requirement for capacitive and resistive touch screen and LC displays. Some exceptional coatings (doped SWNT) showed electro-optical properties approaching those of ITO [6, 51]. In most cases, the *error bars* for these values are not reported, but they could be very large

nanotube networks. Showing the coordinates of sheet resistance and transmittance at 550 nm, the plotted results allow the comparison of the technological requirements for various display applications with the electro-optical properties of the nanotube coatings. At this time, the quality and the FOM of these coatings matches the requirements for resistive, capacitive, and LCD displays. The best coatings, composed of doped carbon nanotubes, show FOM values approaching those of ITO [6, 51]. However, the stability of doping or the possibility of a dedoping during the device operation should be tested before making a definitive conclusion. This suggests that the definition of the FOM can be further developed to include other parameters related to the performance of the coatings for each particular application.

4.6 Resistance Measurements

The accurate assessment of sheet resistance is of fundamental importance for characterization of conductive films. Sheet resistance (Ω/sq.) is the two-dimensional equivalent of 1D linear resistivity (Ω/cm) and 3D bulk resistivity (Ω cm). The resistivity of a wire can be calculated by dividing its resistance by the wire length. The sheet resistivity is a property of a specific film, and it is independent of its geometry. 2D resistivity in applications for nanotube films is related to the in-plane electrical property of the nanotubes, where they are randomly oriented (isotropically). The dimension-independent nature of sheet resistance can be understood using the following example. The resistance of a rectangle with an increasing sample dimension perpendicular to the electrode will increase resistance proportionally. By increasing the sample size along the electrode length, resistance will decrease proportionally. An equal increase of sample size in all directions will therefore have no effect on the value of measured resistance. Thus, the sheet resistance of a square is independent of the square size. A measurement of sheet resistance is also a measurement of sheet resistivity. The units for sheet resistance are ohm per square (Ω/sq.).

There are two options to consider for sample design. The first option is to create a square sample of nanotube film and deposit electrodes on opposite sides of the square. The measured value is the sheet resistance of the nanotube film. Electrodes could be deposited in a multilayer structure composed of 10 A of Ti and 500 A of Au. Ti should be used to improve the contact resistance. One should consider that this is an irreversible procedure, rendering posttreatment of the films to be impossible. The second option involves using spring-loaded point probes placed in contact with the film surface for electrical measurements. The advantage of this procedure is that it does not require any irreversible sample modification.

The electrical properties of the nanotube membranes can be measured using direct and alternating current (DC and AC) techniques, which together provide complementary information and a cross-checking of the accuracy of the measurements.

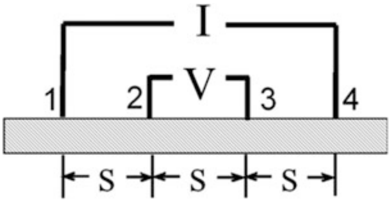


Fig. 4.6 Schematic of a four-probe setup. The probes are separated by equal distances, labeled S. The current is sourced between probes 1 and 4, and the voltage is measured between probes 2 and 3

A source-meter (for instance Keithley 6430) can be used to automatically choose progressively smaller currents for increasingly higher resistances, so that the current will not heat the sample during the measurement process. We estimate that the potential error due to ohmic heating may be in the range of a few degrees, if the minimum current condition is not chosen. We usually use the four-probe technique [40] to estimate the sheet resistance of our samples. The four-probe technique works by contacting four equally spaced point electrodes with the surface sample. The current is sourced between the outer two electrodes, and the voltage is measured across the inner two electrodes. Any contact resistance between the probe tip and the sample is eliminated by nature of the four-probe system; however, an additional correction factor is needed to arrive at the sheet resistance from the measured resistance in a four-probe technique. This factor depends on sample geometry, but values for many different geometries are very well tabulated [41–43].

The elimination of contact resistance in a four-probe setup is easily seen by viewing the system as a series of floating resistances between the current source and sink. In Fig. 4.6, the mutual resistance of point 2 with electrodes 1 and 4 is given by $R_\mathrm{M} = R_{21} - R_{24}$. For both points 2 and 3, the mutual resistance is then $R_\mathrm{M} = R_{21} - R_{24} - R_{31} + R_{34}$. The actual measurement of each resistance would include not only the resistance of the material itself but also the contact resistance as well: $R_\mathrm{ab} = R_\mathrm{sample} + R_\mathrm{contact}$. If the resistance of each contact is the same, the contact resistance will cancel out of the above equation, leaving only the sample resistances.

AC impedance measurements can be conducted using a four- or a two-probe configuration. In our measurements, we usually use the two-probe configuration. At a sufficiently low frequency, AC measurements are expected to agree with the DC measurements, and a relation between two- and four-probe measurements is therefore necessary. Two-probe AC and four-probe AC measurements can be related to one another in two different ways.

The floating potential V at any point away from a current point source on a conductive sheet is given by Uhlir [41]:

$$V_{2\,\mathrm{probe}} = \frac{\rho I}{2\pi} \ln\ r, \tag{4.21}$$

where ρ is the sheet resistivity and r is the distance from the current source.

The sheet resistance of a sample using two point electrodes spaced over distance s, incorporating Ohm's law, is given by:

$$R(s)_{2\,\text{probe}} = \frac{\rho}{2\pi} \ln s \tag{4.22}$$

One can note that subtracting $R(2s)$ from $R(s)$ gives the relation:

$$R(2s)_{2\,\text{probe}} - R(s)_{2\,\text{probe}} = \frac{\rho}{2\pi} \ln 2 \tag{4.23}$$

This is half the four-probe resistance, as given by Smits [42]:

$$R(s)_{4\,\text{probe}} = \frac{\rho}{\pi} \ln 2 \tag{4.24}$$

Therefore, one can measure the two-probe resistance at $1s$ spacing and then again at $2s$ spacing and readily arrive at the measured four-probe resistance. The advantage of this method is that, like a four-probe measurement, the contact resistance is built into the calculations such that it is automatically eliminated upon subtraction. The disadvantage is, of course, that additional measurements are required.

The second method consists of using a conversion factor between the two- and four-probe measurements. It can be shown [40–42] that the floating potentials for a four-probe and two-probe configuration are given by

$$V_{4\text{probe}} = \frac{\rho I}{2\pi} \ln \frac{r_2}{r_1}; \quad V_{2\text{probe}} = \frac{\rho I}{2\pi} \ln r, \tag{4.25}$$

where r_i is the distance from the current source for which the potential is measured. Assuming the two configurations to differ by some multiplicative expression A, one obtains $V_{4\text{probe}} = AV_{2\text{probe}}$; $\frac{\rho I}{2\pi} \ln 2 = A \frac{\rho I}{2\pi} \ln r$, where we have taken $r_2 = 2r_1$, which is the condition for a standard four-probe configuration with the probes equally spaced. We see that for constant tip spacing, s, the factor of conversion, A, between the four-probe and two-probe measurements is constant and independent of source current or sample resistivity.

$$A = \frac{\ln 2}{\ln s}. \tag{4.26}$$

By calculating this constant for the four-probe setup used, it is possible to compare two- and four-probe measurements for all of the samples. While this calculation does not mathematically eliminate the contact resistance, a constant contact resistance would appear into an experimentally determined conversion constant A, still allowing an experimental comparison between measurements.

For the tip spacing we used, $s = 15.9$, measured in units of tip radii [44, 45], since the tip-to-tip spacing was 1.59 mm and the tip radius was estimated as 0.1 mm. Therefore we expect that $A = 0.25$.

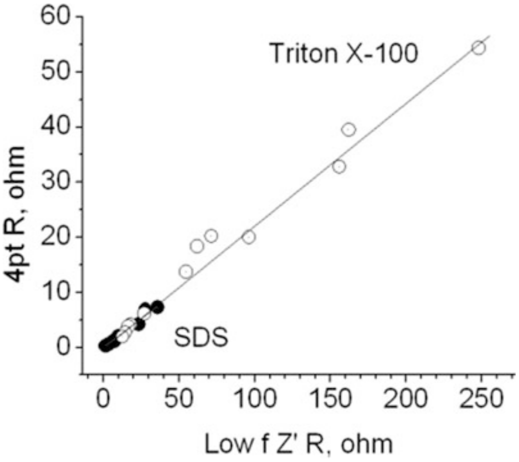


Fig. 4.7 Four-probe resistance of a series of SDS-dispersed (*solid circles*) and Triton X-100 dispersed (*open circles*) SWNT films as a function of two-probe resistance, as measured by low frequency AC measurements. These two values are linearly related by a slope of 0.22

Experimentally, it was found that the four-probe resistance was 0.22 times the two-probe resistance, for a constant distance between probes (Fig. 4.7). This is very close to the expected value of 0.25.

If the contact resistance was a contributor to the two-probe resistance, the experimental value of A would be expected to be smaller than the calculated value, as was observed.

This “contact resistance” is often interpreted as the probe tip being in less than complete contact with the surface and is perhaps better referred to as a “constriction resistance” [44]. Neither the film nor the probe tip is uniformly smooth, so only a certain percentage of the surface area will be in contact. The resistance arises from the current being constricted to certain areas of the probe–film contact, rather than from the tip–film interface itself. Using the experimentally determined value of $A = 0.22$ rather than $A = 0.25$, and substituting it in (4.26) gives an effective tip radius of 0.07 mm rather than 0.1 mm, which implies that only 50% of the tip is in contact with the film.

The comparison between two- and four-probe measurements can then be done by simple multiplication, since this factor of 0.22 was found to be universal among all nanotube films that we measured. Furthermore, the effective tip radius should be used instead of the original tip radius when determining film properties from spreading resistance calculations.

4.7 Impedance Spectroscopy

AC measurements on nanotube networks can shed some light on the electrical transport of individual components of the network, something that would be difficult to obtain from other measurements. By viewing each part of the network as an idealized circuitry component, the contribution of network features to the overall electrical properties can be determined from AC impedance.

Impedance spectroscopy operates by sourcing an alternating potential in the form:

$$V(t) = V_0 \mathrm{e}^{(\mathrm{i}\omega t)}. \tag{4.27}$$

The response signal is a current, shifted in phase from the potential signal:

$$I(t) = I_0 \mathrm{e}^{\mathrm{i}(\omega t - \varphi)}. \tag{4.28}$$

The impedance of the nanotube networks can be expressed in the complex polar form of the impedance magnitude Z_0 and phase angle φ as follows:

$$Z(t) = \frac{V_0 \mathrm{e}^{\mathrm{i}(\omega t)}}{I_0 \mathrm{e}^{\mathrm{i}(\omega t - \varphi)}} = Z_0 \mathrm{e}^{\mathrm{i}\varphi} = Z_0 \cos\varphi + \mathrm{i}Z_0 \sin\varphi. \tag{4.29}$$

For purely resistive (R) behavior, $\varphi = 0$, and the impedance consists entirely of a real term, Z', which is frequency independent. When an inductive and/or capacitive element is introduced, a nonzero value of the phase shift $\varphi \neq 0$ gives rise to the imaginary component of impedance, Z''. At sufficiently low frequency, ω, the imaginary part of the impedance approaches zero, and the real part of the complex impedance, Z', is equivalent to the DC resistance of the system. The direction of the phase shift (the sign of φ) depends on whether the circuit is dominated by inductance or capacitance. A capacitive (C) circuit has negative phase angle, whereas an inductive (L) circuit has positive phase angle [46].

Impedance spectroscopy, along with knowledge of RCL circuitry, therefore allows us to determine which combination of R, L, and C appears in a given network. We know that a capacitor has impedance $Z = 1/\mathrm{i}\omega C$, while an inductor has impedance of $Z = \mathrm{i}\omega L$. The carbon nanotube networks of low density behave as capacitive circuits.

From the superposition principle of Ohm's law, a complex circuit can be reduced to a simpler set of equivalent circuits. Impedance data can be further modeled to deconvolute the circuit elements that the sample consists of.

Modeling of impedance data is done by analysis of the complex plot of $-Z''$ vs. Z', commonly called a Nyquist plot or a Cole–Cole plot. A simple RC circuit in parallel will result in a semicircular Nyquist plot on the complex impedance plane (Fig. 4.8) [46].

From the Cole–Cole plot, the value of the circuit resistance can be obtained from the right-most value, corresponding to $\omega = 0$. The highest point in the arc, the Z'' maximum is the circuit's critical frequency. This critical frequency is evident as the sharp falloff of |Z| in the impedance vs. frequency plot. The capacitance can then be calculated as $C = 1/\omega_c R$, the well-known critical frequency relationship for parallel RC circuits. This measure of capacitance is based on the assumption that the circuit being measured behaves as a simple parallel RC circuit, and the

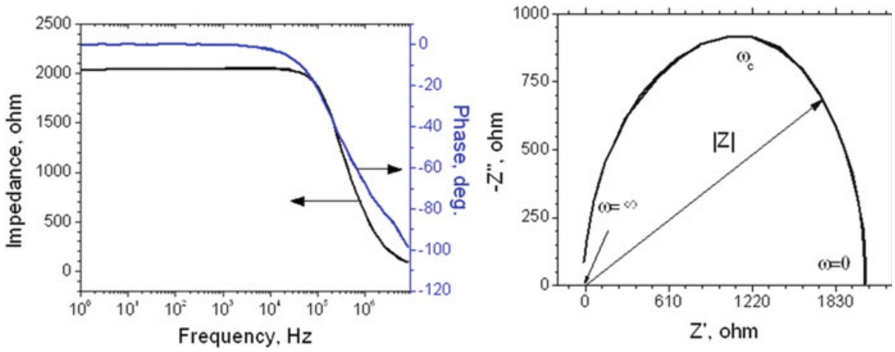


Fig. 4.8 Representative impedance of an SWNT film. |Z| and phase may be shown as a function of frequency, or broken into its real and imaginary components

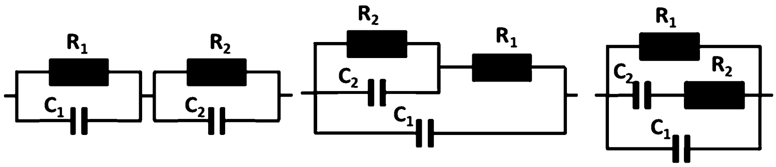


Fig. 4.9 Three different equivalent circuits, which have the same impedance at all frequencies [47]

impedance arc is therefore truly semicircular. However, nanotube networks consist of multiple superimposed arcs [29] and require advanced fitting techniques to determine the proper equivalent circuit to completely describe the electrical response of the sample.

It should be mentioned that multiple equivalent circuits can have the same impedance, yet consist of different elements as shown in Fig. 4.9 [47]. A more extensive analysis, which is beyond the scope of this chapter, is necessary to explain the differences between them (Fig. 4.9).

To model experimental impedance data, one could use a variety of commercially available programs, including the EIS Spectrum Analyser [48] or others. An educated guess at an equivalent circuit should always be made based on the expected structure of the sample, so that the equivalent circuit model proposed will have physical meaning.

4.8 Two-Probe Impedance Results

The impedance response from low density SWNT films are capacitive rather than inductive in nature, as the impedance trends towards zero at high frequency, a characteristic of RC circuits. It was found that impedance spectra from SWNT films

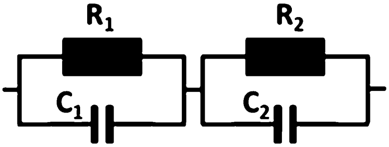


Fig. 4.10 The double-Voigt-element structure fits low density SWNT network impedance well, with one element interpreted as interbundle junctions and the other as the SWNT bundles themselves [29]

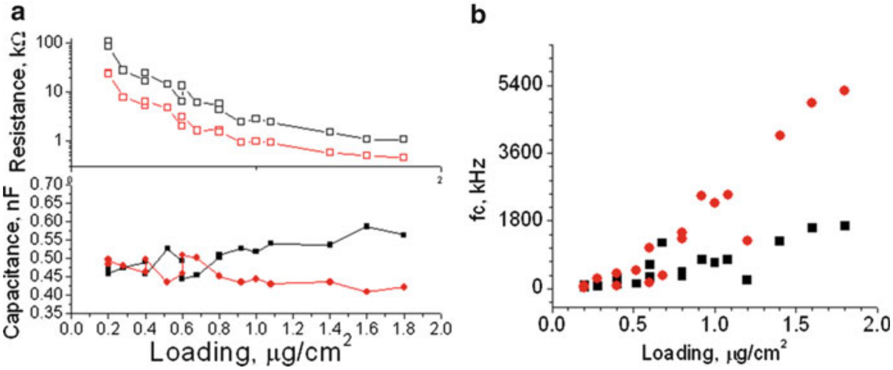


Fig. 4.11 (**a**) The results of modeling the impedance of SDS-dispersed SWNT networks with a double-Voigt-element circuit [29]. The *R* and *C* of nanotube junctions are shown in *black* and *RC* of nanotube bundles are shown in *red*. (**b**) Values of critical frequencies for each Voigt element of SDS-dispersed SWNT films [30]

were best modeled by two Voigt elements in series [29]. A Voigt element is a capacitor and resistor in parallel (see Fig. 4.10). This model makes physical sense, as not only every nanotube but also every junction or defect serves as a resistor, and each will have its own capacitance. It is also the model favored for impedance of bulk polycrystalline systems, with one element being interpreted as the grain boundary and the other as the bulk material [49, 54].

The values of *R* and *C* for these two parallel RC circuits in series were found for films of different thicknesses. The Voigt elements with the lowest *R* value and highest *R* values were plotted in two different curves, as it was assumed that the nanotube bundles and the junctions had similar responses in each film (Fig. 4.11). The corresponding values for capacitance were found to be symmetric, with the average capacitance being a constant (about 0.48 nF) for all loadings well above the percolation threshold for films prepared using Triton X-100 and SDS. If the aforementioned $C = 1/\omega_c R$ relationship were to hold true, this constant network capacitance would mean that the critical frequency and low-frequency conductivity (or four-probe DC conductivity) are linearly related, which we have found to be true.

A similar dependence on film density (loading) has been noted for the critical frequency of each circuit element in our model [30]. The critical frequency of both

Voigt elements increases with loading. For SDS films, each f_c increases at different rates (Fig. 4.11), while for Triton X-100 films, each f_c increases at the same rate (not shown).

The model results indicate that independent of thickness of nanotube networks (loading of nanotubes), the resistance of bundle–bundle junctions is about 3.3 ± 0.3 times higher than the resistance of the bundle, independent of loading, SWNT purity, or dispersant used [29]. This fairly constant value of junction resistance to bundle resistance is expected from the theoretical consideration that the number of junctions increases linearly with the number of bundles [16–18].

4.9 Macroscale Approach to Evaluation of Resistance of Junctions and Bundles in SWNT Percolation Networks by Impedance Spectroscopy

The percolation behavior of the SWNT nanotube networks was investigated for the series of SWNT membranes with increasing loading of nanotubes. Critical percolation threshold and scaling constants were determined for the four-probe DC resistance and low-frequency AC impedance spectroscopy results of SWNT films.

The results indicate that both the AC and DC measurements follow a percolation scaling law, where conductivity, σ can be written in terms of percolation threshold (p_c), the dimension-dependent critical exponent (β) and the equation prefactor (σ_0), which depends on the conductivity of a single SWNT bundle and junction as follows:

$$\sigma = \sigma_0 (p - p_c)^{\beta}. \tag{4.30}$$

The value of critical exponent can take values of 1.33 and 2.0 for two- and three-dimensional conduction. Taking the log of both parts of the percolation equation, we can find that a plot of log(σ) as a function of log($p - p_c$) should result in a straight line. Experimental results for the series of SWNT membranes in these coordinates are shown in Fig. 4.12.

The value of percolation threshold, p_c, was found to be 0.18 ± 0.01 and was dispersant independent for the DC measurements. On the other hand, p_c estimated from the AC measurements was found to be dependent on dispersant and purity. Thus, the lowest value of $p_c = 0.11 \pm 0.01$ was estimated for purified SDS-dispersed SWNTs (see Table 4.2). For the as-grown SWNT films, p_c was found to occur at a loading of 0.25 ± 0.01. The higher value of p_c in the as-grown SWNT can be explained by the contribution of low aspect ratio, carbonaceous impurities. However, purified SWNTs have low amorphous carbon content, and the lower value of p_c for the purified material should have a different cause. One of the potential reasons could be because of the change in the aspect ratio of the SWNTs after purification, which makes them shorter [30]. This is contrary to what should be expected as higher aspect ratio materials normally have lower percolation thresholds. The lower amount of amorphous carbon (with possibly fewer insulating impurities) is a more likely reason for the pc being lower.

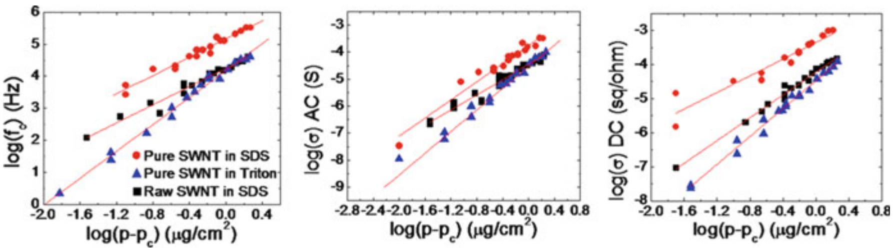


Fig. 4.12 Falloff frequency from real impedance (**a**), low-frequency admittance (**b**), and DC conductivity (**c**), all obey the percolation scaling law, $\sigma = \sigma_0(p - p_c)^\beta$, where σ is conductivity and is proportional to the falloff frequency, p is nanotube loading, with p_c the loading at percolation threshold, and β is the critical exponent. The percolation equation was plotted in a log–log scale. By varying p_c, and choosing the linear fit with the lowest R^2 value, the best fit of the data was found. Values of β are obtained from the slope of the best fit

Table 4.2 A summary of fit parameters attained for the percolation equation by fitting AC and DC experimental results

Dispersant		p_c	β	$\sigma_0 \times 10^4$
Falloff frequency				
Triton X-100	Purified	0.15 ± 0.01	2.11 ± 0.05	1.51 ± 0.12
SDS	Purified	0.12 ± 0.01	1.43 ± 0.07	14.40 ± 1.10
SDS	As-grown	0.25 ± 0.01	1.41 ± 0.06	1.73 ± 0.14
Z′ at low frequency				
Triton X-100	Purified	0.15 ± 0.01	2.01 ± 0.04	0.29 ± 0.02
SDS	Purified	0.11 ± 0.01	1.60 ± 0.06	3.04 ± 0.19
SDS	As-grown	0.25 ± 0.01	1.58 ± 0.07	6.93 ± 0.66
DC resistance				
Triton X-100	Purified	0.17 ± 0.01	2.04 ± 0.05	0.34 ± 0.03
SDS	Purified	0.18 ± 0.01	1.41 ± 0.09	4.07 ± 0.52
SDS	As-grown	0.18 ± 0.01	1.67 ± 0.04	0.60 ± 0.03

In the simplified approach, the SWNT bundles and junctions are assumed to be identical, and the critical exponent is expected to depend on the orientation of the nanotubes only [18]. For random orientation of SWNTs in two dimensions only, the critical exponent takes a value of $\beta = 1.3$, for a three-dimensional networks $\beta = 1.9$. Previous publications have reported $\beta > 1.3$ for nanotube networks [19]. If SWNT bundles and junctions are not identical across the SWNT networks, then β would deviate from the pure instance of 2D conduction [20, 50]. Unlike percolation threshold, the values of the critical exponent, β, were found to range from 1.4 to 2.0 for frequency-dependent and independent measurements, see Table 4.2.

SWNT networks can be described in terms of DC and AC percolation models with the junctions and bundles having a distribution of conductivities, which drives the value of the critical exponent, β, higher than expected for a 2D network. The contributions of junctions and bundles to the macroscopic impedance of a nanotube network can be separated by modeling AC data with a double-Voigt-element model as described earlier. This approach will provide essential information for optimization of nanotube-based transparent conductive films.

4.10 Effect of Doping on the Frequency-Dependent Impedance

For doped films of similar thickness and method of preparation, the relation between the critical frequency and low frequency (DC) conductivity also exhibits a linear dependence. Figure 4.13a shows that the critical frequency depends on the degree of doping and can be determined based on choice of dopant and method of doping. Figure 4.13b on the right displays the effect of doping type on the Nyquist plots of SWNT films of 1.4 μg cm^{-2} density. This means that, just like conductivity can be selectively determined by doping, so can f_c. In-situ control of f_c would also be possible by monitoring the DC conductivity, which is experimentally much more easily determined in real time, as a sweep of many frequencies is not necessary.

It is notable that the shape of the arc in the Cole–Cole plot does not change upon doping, only the size of the arc changes. This means that the model applied to undoped films is still valid for doped films.

4.11 Effect of Nanotube Length on Transparency and Conductivity of CNT Electrodes

The aspect ratio of sticks comprising a conductive network is inversely proportional to the percolation threshold, the minimum amount of a material required for conduction. The high aspect ratio of carbon nanotubes results in the percolation threshold for these materials at less than 1% by volume or area. In fact, for the series of films described in Fig. 4.11 and Table 4.2, the percolation threshold was 0.18 on average [30].

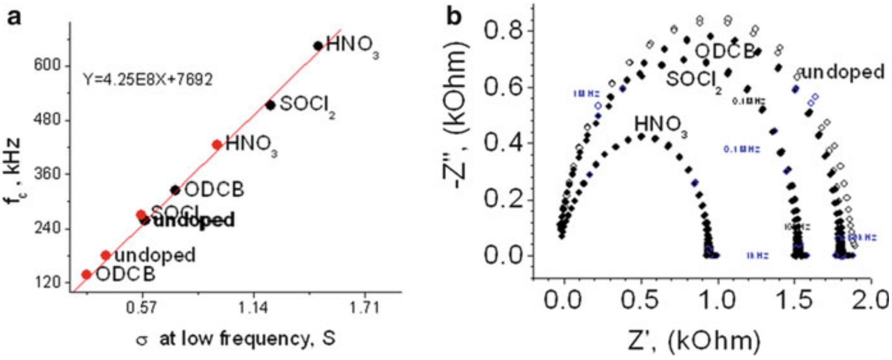


Fig. 4.13 (**a**) The dependence of critical frequency on low frequency admittance (DC conductivity). Vapor-doped specimens are shown in *red*, liquid-doped in *black*. (**b**) Nyquist plot of 1.4 μg cm^{-2} SWNT films exposed to vapors of dopants, which include orthodichloro benzene, nitric acid, and thionyl chloride [30]

Most of the SWNT networks are composed of nanotubes with some broad distribution of the aspect ratios, which in most instances, is difficult to characterize. At the same time, most models are done for a fixed aspect ratio, thus making it hard to relate the modeling to the experimental result.

The effect of nanotube aspect ratio on the transparency and conductivity of nanotube networks is also difficult to confirm experimentally as it requires separation of nanotubes by length. Assuming that the diameter of nanotubes for a particular synthesis condition is constant, nanotubes with different aspect ratio can be produced by dispersion of nanotubes through the sonication followed by dense liquid centrifugation [7]. Other techniques have been shown to be effective in fractionalization of nanotubes by length including gel electrophoresis [8] and size exclusion chromatography (SEC). For instance, using gel electrophoresis and column chromatography on cup horn sonicated nanotubes, Strano et al. resolved fractions of nanotubes with average length between 92 and 435 nm [8]. One should be thorough in analyzing the fractions of nanotubes obtained, as instead of fractionalization, chiral separation can be obtained along with length fractionalization [9, 52]. Strano also reported a concomitant enrichment of large diameter nanotubes in fractions with shorter nanotubes.

In another effort, iodixanol (5,5′-[(2-hydroxy-1-3 propanediyl)-bis(acetylamino)] bis [*N*,*N*′-bis(2,3 dihydroxylpropyl-2,4,6-triiodo-1,3-benzenecarboxamide)]) has been used to generate various density solutions and achieve length-fractioned samples of SWCNTs (laser, CoMoCat, and HiPco) [7, 10]. To achieve density modification, a surfactant (sodium deoxycholate) solution was mixed with iodixanol and layered in the centrifuge tube. The length-fractioned nanotubes showed no evidence of chirality specific separation as evident from the solution absorbance spectra [7]. Simien et al. used length fractionation to investigate the effect of nanotube length on the optical and electrical properties of SWNT films [11]. The films composed of 130, 210, and 820 nm long and mixed lengths of SWNTs (CoMoCat) were tested. It was found that the absorbance of nanotubes scales linearly with their density and the changes in conductivity can be quantitatively described by the generalized effective medium approximation.

The resulting percolation curves are shown in Fig. 4.14 (summarized from Figs. 4 and 5 in [11]). This graph shows the clear advantage of using length-separated tubes, compared to SWNT of mixed length. A detailed analysis of the properties of SWNT around the percolation threshold indicated that the networks can be described as nearly 2D networks for long SWNTs and as 3D networks for short ones. The authors also demonstrated that experimental results correlate well with percolation theory and the conductivity percolation threshold (p_c) varies with the aspect ratio L as, $p_c = 1/L$. It is interesting to note that the most drastic effects were observed around the percolation threshold of films made from 0.018 μg cm^{-2} (820-nm long tubes). The networks prepared from separated SWNTs show similar conductivities for loading >0.2 μg cm^{-2}. The distinction between mixed and separated nanotubes disappears for SWNT loading exceeding 1.5 μg cm^{-2}.

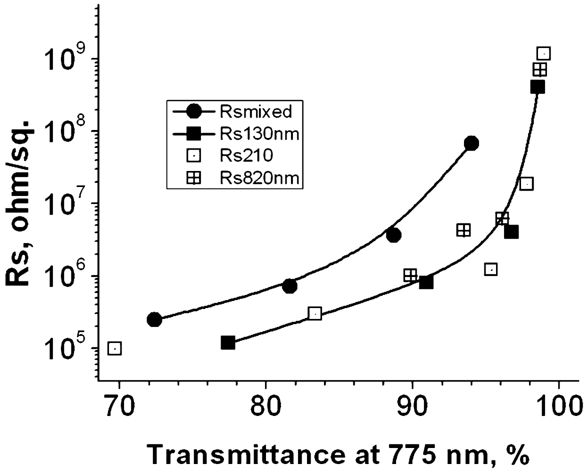


Fig. 4.14 Effect of nanotube length fractionation on the sheet resistance and the transmittance of their conducting networks. Data were extracted from Figs. 4 and 5 in [11]. TCE prepared from separated nanotubes show better quality factor compared to those built from mixed length tubes. The 775-nm wavelength was selected because at 550 nm CoMoCat SWNT exhibit higher absorbance due to the $E_{11}{}^{S}$ transition

4.12 Effect of Semiconducting and Metallic Nanotubes on the Transparency and Sheet Resistance of TCCs

As produced CNTs contain about 75% semiconducting nanotubes, which do not contribute to the conductivity but reduce the transmittance of the networks. Therefore, one would expect that TCC from the networks of pure metallic nanotubes would demonstrate best FOM. We should remember that a linear dependence of the optical properties with the amount of nanotubes exists only for the absorbance. Thus, one might expect that the network of unsorted nanotubes with sheet resistance of 10 Ω/sq. and 66% of transmittance may be possible to turn into films with 90% transmittance with the same resistance if only metallic nanotubes were used. However, most of the nanotube networks show sheet resistance above 1,000 Ω/sq., which is a factor of 10 larger.

Several approaches have been identified for sorting nanotubes into metallic and semiconducting types for postprocessing, including gel-separation [12], density gradient ultracentrifugation (DGU) [13], dielectrophoresis, chemical selection, electrical breakdown, chromatography, and selective growth; yet, the most popular method for large-scale production of metallic and semiconducting nanotubes is DGU and has been recently realized on the commercial scale by Nanointegris.

Interesting comparisons of all-metallic and unsorted (a mixture of metallic and semiconducting) SWNT were made by the Yang, Blackburn, and Hersam groups [14–16]. They demonstrated that semiconducting nanotubes and unsorted tubes show higher sheet resistances compared to networks composed of pure metallic nanotubes. The network composed of all-metallic nanotubes shows 150 Ω/sq. and 80% transmittance, while similar networks for unsorted nanotubes show 200 kΩ/sq. and 82% transmittance (see Fig. 4.14).

The absolute improvement in sheet resistance is 1,000 times [14]. Laser tubes showed better FOM (lowest *R* and highest *T*) compared to HIPCO and arc tubes [14]. The acidic and $SOCl_2$ doping of semiconducting and metallic networks improves their sheet resistance by a factor of 10 and 4, respectively, with almost no effect on their transmittance [15]. Green and Hersam showed that all metallic nanotubes improve the conductivity of networks by a factor 5.6, with the best network showing ~90% transmittance and sheet resistance of 500 Ω/sq. [17]. According to Miyata et al., metallic nanotube networks show relatively small difference in resistance compared to unsorted nanotubes, and a slightly larger improvement of resistance for unsorted tubes upon doping with sulfuric acid. In both cases, the transmittance of nanotubes was in the 40–48% range and showed 2–5% improvement in doped samples [18].

Tyler et al. showed that the sheet resistance of nanotube electrodes can be tuned by varying the amount of metallic tubes and the doping level (nitric acid or PEDOT:PSS) [16]. They demonstrated that TCCs deposited from 99.9% semiconducting nanotubes could get lower sheet resistance (188 Ω/sq.) compared to those produced from 99.9% metallic nanotubes (411 Ω/sq.) with close transmittance values (88–90%). However, the performance in OPV is the opposite, with metallic CNT TCCs showing higher open-circuit voltage, short-circuit current, and fill factor compared to that for doped semiconducting nanotubes. The short-circuit current showed the most drastic difference, 300 times higher for doped metallic CNT TCE [16].

We have previously noted that nanotube networks could behave differently even if they are assembled from the same type of nanotubes [29, 30]. The differences may come from the purity, the bundle size, and/or the aspect ratio, all of which can result in significant changes in the electro-optical properties. Without identifying all of the important parameters of the networks, it would be hard to make side-by-side comparison with other published results. Thus, absolute changes in the measured sheet resistance of semiconducting nanotubes upon doping may be a result of smaller aspect ratio networks with a larger number of junctions for the metallic tubes rather than nanotube type. This is one of the reasons why it is suggested that instead of reporting an absolute improvement on the sheet resistance or transmittance, a comparison of the appropriate FOM should be used. Figure 4.14 summarizes the different trends just described.

4.13 The Effect of the Number of Walls on the Quality of Carbon Nanotube-Based Transparent Conducting Electrodes

Possible damage (introduction of sidewall defects) of single-wall carbon nanotubes during their dispersion under sonication may reduce the quality of the nanotubes. One solution is to use nanotubes with larger number of walls as material for TCEs.

In this case, the outer wall of the nanotube can be sacrificed/damaged to enable dispersion, but the inner wall would stay undamaged for charge carrier transport. A comparison of transparent and conductive electrodes composed of single-wall nanotubes with a diameter of 0.6–0.9 nm, double-wall carbon nanotubes with a diameter of 1.5–2.7 nm, and multiwall nanotubes with diameter of 16–35 nm was reported by Li et al. [19]. The small variation in the diameter of the nanotubes was compensated by length, resulting in relatively similar aspect ratio for the series of samples. For the same transmittance, higher conductivity was measured for $SOCl_2$-doped SWNTs followed by double-wall nanotubes, MWNTs, and undoped SWNTs [19]. The authors rationalized the results by ascribing them to better conductivity of individual MWNTs as compared to SWNTs. It is interesting to note that the effect of $SOCl_2$ doping is more pronounced (factor of 10 improvement in conductivity) for thicker coatings, and much smaller for thin coatings. Green and Hersam used GDC to separate DWNTs from a mixture of as produced SWNTs and DWNTs and measured their optical and electrical properties in solution and as coatings [20]. They reported that for the same transmittance, enriched DWNTs doped with $SOCl_2$ make 42% more conductive networks compared to those of undoped SWNT. Pristine, undoped, DWNTs coatings are a factor of 2.4 better conducting compared to SWNTs [20].

Li et al. reported on TCCs prepared from pristine DWNTs (Xin Nano Materials) [21]. The purification of nanotubes was done by air oxidation, followed by the hydrochloric acid wash to remove catalyst and oxide nanoparticles. Purified DWNTs (p-DWNTs) were obtained by treating purified tubes with a mixture of sulfuric and nitric acids for functionalization (f-DWNT). The electro-optical properties of nanotube films prepared from DWNTs (pristine-, p-, and f-DWNT) are shown in Fig. 4.15. Purification and functionalization of DWNTs led to better quality of material, mostly due to removal of amorphous carbon. The best f-DWNT demonstrated sheet resistance of about 1.1 kΩ/sq. and transmittance of 90% [21]. MWNTs are expected to exhibit higher absorbance for the same density of nanotubes as compared to the SWNT, DWNT, or FWNTs. Earlier reports of similar electro-optical properties of DWNT and MWNT TCCs may stem from some morphological differences (aspect ratio) or possibly a mixed character of CVD-grown DWNT [22]. Considering the much lower cost of MWNTs compared to other nanotubes, the possibility of creating an inexpensive conductive coating prompted extensive research. The electro-optical properties of MWNT TCCs, summarized by Kaempgen [23], showed a percolation-like behavior observed for coatings with the transmittance of 85–90%.

Castro et al. reported on processing 10 nm MWNT (from Nanocyl) into conductive coatings without additional purification [24, 53]. Ko prepared TCCs from 40 nm diameter and 1–2 μm long MWNTs [25]. The TCCs made from the solvent containing drops of MWNT suspension produced a macroscopic structure of overlapping rings with a sheet resistance of 300 Ω/sq. and transmittance at 550 nm of about 80% [25]. The electro-optical properties of TCCs composed from nanotubes with different number of walls are summarized on Fig. 4.15. As anticipated, the tubes with a smaller number of walls exhibit better properties, and highly enriched DWNT take the lead in this group [20].

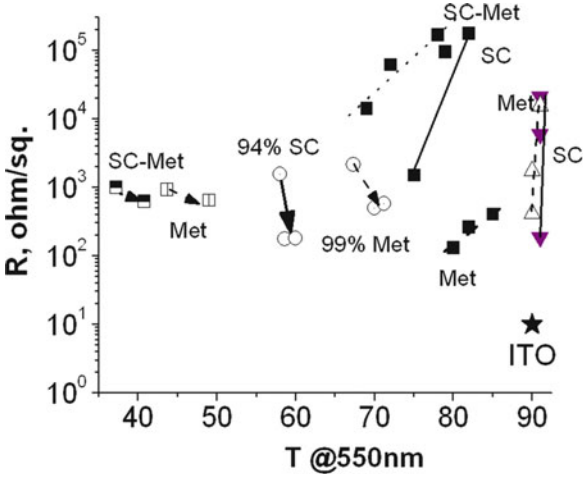


Fig. 4.15 Summary of the effect of nanotube character (semiconducting or metallic) on transmittance and sheet resistance of a nanotube transparent conducting electrodes. *Circles*—data adapted from [14] show the effect of film thickness (1, 2, and 3 layers) on the *R–T* parameters of pure (metallic and semiconducting) and mixed films. Data shown in *open circles* were adapted from the [15]. The effects of doping on sheet resistance shown in *triangles* were obtained from [16]. Data shown as *semi-solid squares* were taken from [18]. *Broken lines* connect data for all metallic nanotube TCEs from the same reference. The TCC with the best properties shown above was prepared from semiconducting nanotube doped with nitric acid [16]

4.14 The Effect of Metal Nanoparticles on the Quality of Carbon Nanotube-Based Transparent Conducting Electrodes

While doping with metal nanoparticles seemed to be attractive for reduction of tube-to-tube resistance and sidewall defects, one we should be aware that such p-doping may also downshift the SWNT work function up to 0.42 eV by a strong charge transfer from the nanotubes to $AuCl_3$ [26]. Also, as with any doping, we should consider the stability of the metal nanoparticles, as well as the potential instability of ligand doping [27]. At elevated temperatures, Au nanoparticles have high mobility on carbon due to the low melting point of nanoparticles and the noncovalent character of their coordination. This will lead to growth in nanoparticle size, leading to increase of sheet resistance and reduction in transmittance due to large contribution of light scattering. Anion doping may not be stable either, even for stable ions like Cl^- [27]. Park et al. reported on decoration of purified and functionalized bamboo MWNT (Iljin Nanotechnology Inc.) [50]. Functionalization of MWNTs was done by the reaction with 1-butyl-3-methylimidazolium tetrafluoroborate (BMITB) in the solid phase at room temperature. MWNTs were then washed with acetonitrile and deionized water to remove excess of BMITB and dried under vacuum. Au decoration of functionalized MWNTs was achieved

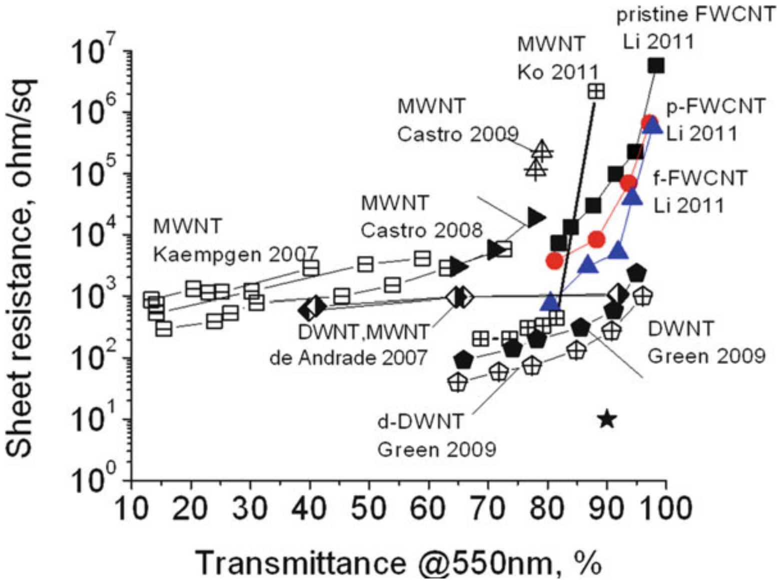


Fig. 4.16 Electro-optical properties of double-wall, few-wall, and multiwall nanotubes presented in coordinates of sheet resistance and transmittance. Although not all reports followed properties of nanotube networks through the percolation, a clear tend can be seen. The smaller number of walls, the lower is the position of percolation curve on the plot. The electro-optical properties of ITO are shown with *solid star symbol* (10 Ω/sq., 90 % *T*). The best FOM is shown by enriched d-DWNTs [20]

through a reaction with $HAuCl_4{\cdot}3H_2O$ in the solution of Et/H_2O mixed in 1:1 ratio. The mixture was briefly sonicated using 240 W at 20 kHz to initiate Au nanoparticles. The membranes of Au-decorated functionalized MWNT (Au-f-MWNT) were prepared by a filtration method, followed by dissolution of the membrane and deposition of TCC on a substrate [50]. Park et al. report more than twofold lower sheet resistance for Au-f-MWNT (Au nanoparticles with diameters of 10.3 ± 1.5 nm) compared to original MWNTs due to the well-interconnected three-dimensional structure, which incorporates Au nanoparticles. For Au-f-MWNT films, with a transmittance of 88.3%, the sheet resistance was found to be about 42.5 kΩ/sq. (see Fig. 4.16).

Yang et al. compared acidification of nanotube networks with the effect of Au-decoration of purified DWNT and SWNTs (HiPco and Unidyme) on the electro-optical properties of the TCCs produced from these materials [28]. The nanotubes were first dispersed in a mixture of water and propanol, with the help of Nafion, which stimulates electrostatic and steric stabilization of the nanotubes, leading to a stable dispersion of the nanotubes in the solvent mixture. For Au-decoration, nanotube films were immersed in a 1 mM gold salt solution in $HAuCl_4{\cdot}3H_2O$ in 50 vol. % for 10 min. For acidification, the nanotube films were soaked in concentrated HNO_3 (60% pure), followed by washing in deionized water.

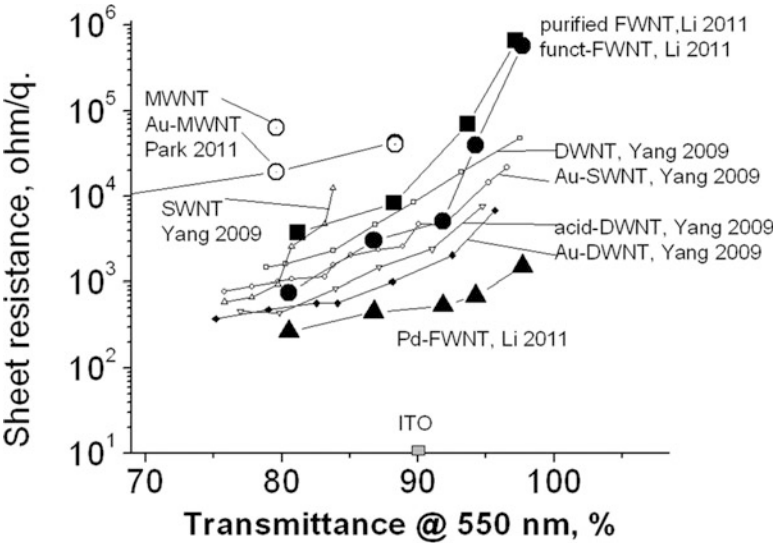


Fig. 4.17 Effect of metal decoration on the electro-optical properties of carbon nanotube TCCs, as compared to uncoated MWNT, DWNT, and f-DWNT

Au decorated DWNT (with a diameter of 0.9–1.2 nm and length of 10–30 μm) showed a better quality than the rest of the nanotubes (1 kΩ/sq. and 90% transmittance). The nanotubes are arranged in the following order based on the quality of the properties of the transparent conductive films: Au-DWNTs > acid-DWNTs> Au-SWNTs > pristine DWNTs > pristine SWNTs, which are summarized in Fig. 4.17.

Li et al. reported that decoration of a few wall carbon nanotubes by Pd also reduces the resistance of nanotube networks [21]. Pd decoration was achieved by reduction of palladate salts on the surface of nanotube networks with hydrogen at 500 °C for 30 min. This process produces homogeneous nanoparticles of Pd, with average diameter of 5 nm. The nanotube ink was formed by dispersion of 30 mg Pd-FWNT in polyvinylpyrrolidone in 120 mL of ethanol, followed by centrifugation and collection of the top layer of the solution. A two order of magnitude reduction in sheet resistance of Pd-FWNT was achieved, with the best TCC demonstrating about 274 Ω/sq. and 81.65% transmittance, as shown in Fig. 4.17.

4.15 Concluding Remarks

This chapter described the methodology for properly assessing the necessary optical and electrical properties needed to use carbon nanotube assemblies as transparent conducting coatings, with some special requirements for some specific applications. The first part of the chapter focused on defining the FOM, a quantity that can be used to determine the best combination of the optical transmission and the sheet conductance necessary to maximize the film's performance. Various FOMs have been reviewed and used to demonstrate their validity using results available from the vast literature in this field. In the second part, a description of how to characterize the electrical properties of the nanotube thin films was presented. The procedure for how to relate four-probe sheet resistance measurements and two-probe AC measurements was also included. In the last part of the chapter, data showing the effects of varying the nanotube characteristics (SWNT, DWNT, and MWNT) was summarized in several useful sheet resistances versus transmission graphs, where the effects of nanotube length as well as fractions of metallic and semiconducting nanotubes and the effect of doping were considered.

Acknowledgments This work was conducted at the Center for Nanophase Materials Sciences, which is sponsored at Oak Ridge National Laboratory by the Office of Basic Energy Sciences, U.S. Department of Energy. We acknowledge financial support by the Solar Energy Technologies Program of the Office of Energy Efficiency and Renewable Energy, U.S. Department of Energy.

References

1. US Geological Survey, *Mineral commodity summaries 2011* (US Geological Survey, Reston, VA, 2011), p. 198
2. C.A. DiFrancesco, M.W. George, Jr.J.F. Carlin et al., ed. by U.S. Department of the Interior (U.S. Geological Survey, Reston, VA, 2010). Indium Statistics
3. L. Hu, D.S. Hecht, G. Gruner, Nano Lett. **4**(12), 2513 (2004)
4. B. Ruzicka, L. Degiorgi, R. Gaal et al., Phys. Rev. B **61**(4), R2468 (2000)
5. A. Pekker, K. Kamaras, J. Appl. Phys. **108**(5), 054318 (2010)
6. D.S. Hecht, A.M. Heintz, R. Lee et al., Nanotechnology **22**(16), 5 (2011)
7. J.A. Fagan, M.L. Becker, J. Chun et al., Adv. Mater. **20**(9), 1609 (2008)
8. D.A. Heller, R.M. Mayrhofer, S. Baik et al., J. Am. Chem. Soc. **126**(44), 14567 (2004)
9. C.A. Dyke, M.P. Stewart, J.M. Tour, J. Am. Chem. Soc. **127**(12), 4497 (2005)
10. S.M. Tabakman, K. Welsher, G. Hong et al., J. Phys. Chem. C **114**(46), 19569 (2010)
11. D. Simien, J.A. Fagan, W. Luo et al., ACS Nano **2**(9), 1879 (2008)
12. T. Tanaka, H. Jin, Y. Miyata et al., Nano Lett. **9**(4), 1497 (2009)
13. Y. Feng, Y. Miyata, K. Matsuishi et al., J. Phys. Chem. C **115**(5), 1752 (2011)
14. A. Rahy, P. Bajaj, I.H. Musselman et al., Appl. Surf. Sci. **255**(15), 7084 (2009)
15. J.L. Blackburn, T.M. Barnes, M.C. Beard et al., ACS Nano **2**(6), 1266 (2008)
16. T.P. Tyler, R.E. Brock, H.J. Karmel et al., Adv. Energy Mater. **1**(5), 701 (2011)
17. A.A. Green, M.C. Hersam, Nano Lett. **8**(5), 1417 (2008)
18. Y. Miyata, K. Yanagi, Y. Maniwa et al., J. Phys. Chem. C **112**(10), 3591 (2008)
19. Z. Li, Appl. Phys. Lett. **91**(5), 053115 (2007)
20. A.A. Green, M.C. Hersam, Nat. Nano. **4**(1), 64 (2009)
21. Y.-A. Li, N.-H. Tai, S.-K. Chen et al., ACS Nano. **5**(8), 6500 (2011)
22. M. Jung de Andrade, D.L. Márcio, V. Skákalová et al., Physica Status Solidi Rapid Res. Lett. **1**(5), 178 (2007)
23. M. Kaempgen, G.S. Duesberg, S. Roth, Appl. Surf. Sci. **252**(2), 425 (2005)
24. M. Castro, N. Al-Dahoudi, P. Oliveira et al., J. Nanopart. Res. **11**(4), 801 (2009)
25. W.-Y. Ko, J.-W. Su, C.-H. Guo et al., Thin Solid Films **519**(22), 7717 (2011)

26. K.K. Kim, J.J. Bae, H.K. Park et al., J. Am. Chem. Soc. **130**(38), 12757 (2008)
27. S.M. Kim, K.K. Kim, Y.W. Jo et al., ACS Nano **5**(2), 1236 (2011)
28. S.B. Yang, B.-S. Kong, J. Geng et al., J. Phys. Chem. C **113**(31), 13658 (2009)
29. M.P. Garrett, I.N. Ivanov, R.A. Gerhardt et al., Appl. Phys. Lett. **97**(16), 163105 (2010)
30. M.P. Garrett, Ph.D. Dissertation, University of Tennessee, Knoxville, 2009
31. R. Baetens, B.P. Jelle, A. Gustavsen, Solar Energy Mater. Solar Cells **94**(2), 87 (2010)
32. ASTM International, 2008
33. C.G. Granqvist, Solar Energy Mater. Solar Cells **91**(17), 1529 (2007)
34. T.Y. Crowell, *The Science of Color* (Optical Society of America, New York, 1953)
35. G. Haacke, Annu. Rev. Mater. Sci. **7**(1), 73 (1977)
36. G. Haacke, J. Appl. Phys. **47**(9), 4086 (1976)
37. R.G. Gordon, MRS Bull. **25**, 52 (2000)
38. R.E. Glover, M. Tinkham, Phys. Rev. **108**(2), 243 (1957)
39. B.S. Shim, J. Zhu, E. Jan et al., ACS Nano **4**(7), 3725 (2010)
40. L.B. Valdes, Proc. IRE **42**(2), 420 (1954)
41. A. Uhlir, Bell Syst. Techn. J. 105 (1955)
42. F.M. Smits, Bell Syst. Techn. J. 710 (1958)
43. V.S.K.G. Kelekanjeri, R.A. Gerhardt, Meas. Sci. Technol. **19**(2), 025701 (2008)
44. R. Holm, *Electric Contacts: Theory and Application*, 4th edn. (Springer, New York, 1967)
45. J. Albers, H.L. Berkowitz, J. Electrochem. Soc. Solid State Sci. Technol. **131**(2), 392 (1984)
46. R.A. Gerhardt, in *Encyclopedia of Condensed Matter Physics*, ed. by G. Bassani, G.L. Liedl, P. Wyder (Elsevier Press, Oxford, 2005), p. 350
47. I.M. Novosel'skii, N.N. Gudina, Y.I. Fetistov, Soviet Elektrokhimiya **8**, 565 (1972)
48. A.S. Bondarenko, G.A. Ragoisha, EIS Spectrum Analyser, http://www.abc.chemistry.bsu.by/vi/analyser/
49. J.R. Macdonald, *Impedance Spectroscopy* (Wiley, New York, 1987)
50. H.S. Park, J.-S. Kim, B.G. Choi et al., Carbon **48**(5), 1325 (2010)
51. D.S. Hecht, L. Hu, G. Irvin, Adv. Mater. **23**(13), 1482 (2010)
52. D. Chattopadhyay, I. Galeska, F. Papadimitrakopoulos, J. Am. Chem. Soc. **125**(11), 3370 (2003)
53. M.R.S. Castro, H.K. Schmidt, Mater. Chem. Phys. **111**(2–3), 317 (2008)
54. R. Gerhardt, A.S. Nowick, J. Am. Ceram. Soc. **69**(9), 641 (1986)

Chapter 5
Silicon Electroplating for Low Cost Solar Cells and Thin Film Transistors

Dominic F. Gervasio and Olgierd Palusinski

Abstract Silicon electroplating offers a low-cost method for the production of high-performance low-cost silicon solar cells that can be used in small portables and large-scale applications, like the grid. Silicon remains the semiconductor of choice because silicon has the best combination of efficiency, cost, durability, and availability. Silicon photovoltaic (PV) devices are likely to dominate the market for a long time. Silicon solar cells have reasonable efficiency (up to 15%), cost (as low as $2/peak watt), and excellent reliability (losing less than 1% power output per year over 25 years), and since silica is abundant, silicon depletion is not a worry. Although silicon is the best photovoltaic option and has the largest market share, it is still too costly to provide the majority of grid power. Cost remains a major barrier to further market penetration, because current thin film semiconducting silicon preparation uses high-temperature (750–1,000°C) deposition processes, such as chemical vapor deposition (CVD), which require high levels of electrical power and energy and convert only 10% of the silane feed to useful silicon. Clearly silicon PV manufacturers need to increase efficiency and lower wastes and cost. Silicon electrodeposition offers an effective alternative to CVD for making silicon devices with substantially reduced processing costs so that solar photovoltaics can be cost competitive with the typical cost for installing new electrical power generators in the grid. Using silicon electrodeposition as the silicon processing in the manufacture of a variety of semiconductor applications is reviewed. A practical way of electroplating silicon from silicon salts dissolved in ionic liquids is discussed with early results and prospects.

D.F. Gervasio (✉)
Department of Chemical and Environmental Engineering, University of Arizona,
Tucson, AZ 85721, USA
e-mail: gervasio@email.arizona.edu

O. Palusinski
Department of Electrical and Computer Engineering, University of Arizona, Tucson,
AZ 85721, USA

A. Korkin and D.J. Lockwood (eds.), *Nanoscale Applications for Information and Energy Systems*, Nanostructure Science and Technology,
DOI 10.1007/978-1-4614-5016-0_5, © Springer Science+Business Media New York 2013

5.1 Introduction

Bob Dylan sang "…the times they are a changing" to urge social reform. These words could just as well lead the call for the reform of industrial practices. In the twentieth century, the growth of industry was literally fueled by oil and was so successful that it has grown itself out of business as usual. The manufacturing practices and energy technology that spurred industrialization over the last century are becoming obsolete. The "energy crisis" of the 1970s leading to a threefold rise in oil prices was really an "oil-distribution crisis" and was largely forgotten when oil prices fell. However, oil prices are rising again because after 50 years of unchecked consumption there is a real "energy-supply crisis." And 50 more years of the same invites disaster.

Clean efficient manufacturing practices are not just desirable, but necessary to preserve acceptable water and air quality as well as to conserve fuel, because petroleum supplies are rapidly being depleted. New sources of energy are needed. The purpose of this tutorial is to illustrate alternatives to conventional manufacturing processes and energy technology. By developing clean and efficient alternatives we can continue the quality of life that we have come to enjoy.

The technologies presented here are by no means the only or best ones, but are offered as examples of reasonable alternatives that can be achieved with reasonable effort. The prime example of a clean efficient alternative industrial practice which is offered in this chapter is using energy efficient electroplating to make silicon in place of using the dirty and energy-intensive CVD methods which have traditionally been used for the production of semiconducting silicon devices. Electroplating silicon can lower the cost of photovoltaic devices so that solar energy can become a large-scale sustainable energy source instead of petroleum fuel. In summary, successful electroplating of high purity silicon would allow photovoltaic manufacturing and deployment to be more accessible, because electroplating processing is far less polluting, energy intensive, and expensive than traditional CVD methods.

5.2 Thin Film Semiconducting Devices: The Big Picture

To address the changing demands for global energy, it is wise to begin *now* to develop new approaches for clean and efficient manufacturing practices and sustainable power sources. One approach that addresses both of these issues is the electrochemical deposition (electroplating) of semiconductors, like silicon. Electrodeposition is a clean, effective, inexpensive, and proven industrial processing method. A large-scale industrial use of electrodeposition is for plating copper conduction lines in printed circuits. Now electrodeposition can offer an alternative to conventional silicon processing for making a variety of desirable devices that operate using a thin layer of semiconducting silicon. These devices include:

1. Thin film transistors (TFTs)
2. Photovoltaics (PVs)
3. Hybrid photovoltaic and capacitor device

Device 3 is a power source device and an energy storage device in one hybrid device called a renewable hybrid power source (RHPS).

For electronic devices like TFT circuits, cost is not an issue. Electronics like computers, radios, televisions, music players, etc., are premium products whose prices can cover the high cost of producing an ultra pure silicon wafer by CVD methods. However, power and energy are relatively low-cost commodities, so cost is an important and limiting consideration when producing silicon to be used in photovoltaics for production of solar power. Presently, photovoltaic power costs on the order of $5,000 per kW and is projected to drop to about $4,000 per kW by 2015 [1]. *This 20% cost reduction for silicon photovoltaics made by conventional methods is simply not enough*. If photovoltaic power sources are to be applied to grid power then cost must drop by orders not fractions. The present capital investment for grid power is only $125 per kW.

There are numerous advanced compound semiconducting materials (such as CdTe, Cu(In,Ga)Se_2, etc.) under development to improve solar to electrical conversion efficiency. These advanced semiconductors are more expensive than silicon, and even if in mass production these lead to percentages of cost reduction, this is not the desired orders of magnitude of cost reduction needed to make solar photovoltaics cost competitive for grid application. Although advanced materials are attractive from an efficiency standpoint, it is clearly cost reduction, which is not just desirable, but which is an absolutely critical, if solar to electrical conversion is to enter the mainstream of electrical power supply. Electroplating is a way to dramatically drop the costs of the production of silicon and possibly the advanced photovoltaic materials as well. The implication and suggestion is that new low cost processing is the key. Advanced processing of known materials—not the discovery of advanced semiconductors—seems to be the more critical need for bringing large-scale photovoltaic power to market.

There are essentially three technical considerations that favor electroplating as a process for the production of semiconducting silicon. Electroplating is

1. Inexpensive, requiring little energy because it is a high efficiency process.
2. Clean, yielding no VOCs, noxious waste, or fumes.
3. Allows the formation of *complex Si structures*.

The third consideration allows formation of nanowires and vertically oriented nano-Schottky diodes structures (see Fig. 5.1) which are virtually impossible to make using conventional Si processing.

A tool for forming silicon into complex nanostructures is intriguing. After photon absorption induces charge separation in the silicon, there is less of a chance the charges will recombine if the photoelectron and hole have a short path to the current collectors.

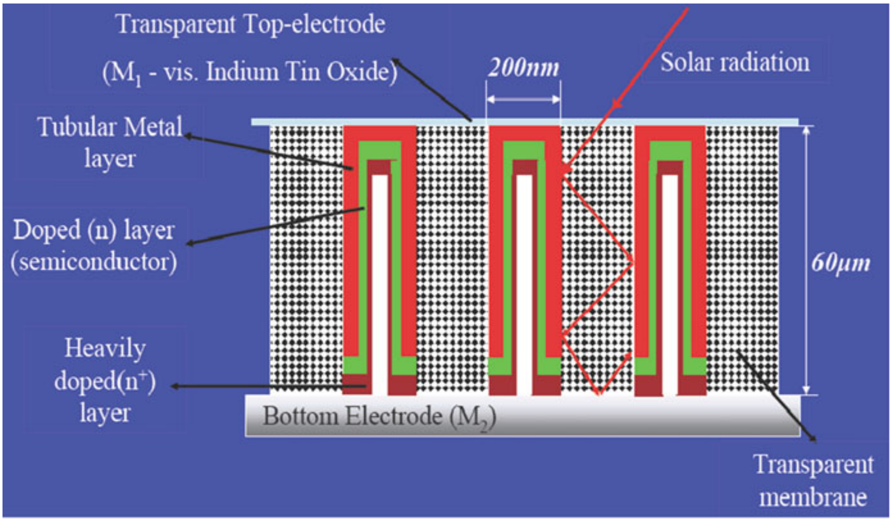

Transparent Top-electrode
(M$_1$ - vis. Indium Tin Oxide)
200nm
Solar radiation
Tubular Metal layer
Doped (n) layer (semiconductor)
60μm
Heavily doped(n$^+$) layer
Bottom Electrode (M$_2$)
Transparent membrane

Fig. 5.1 Schematic diagram of a vertically oriented nano-Schottky diode for efficient radiation capture

A Schottky diode made by electroplating silicon on a metal is one of the simpler ways to make a photovoltaic. Forming structures like the Schottky-diode array shown in Fig. 5.1 is a viable way to make efficient photovoltaics which could help meet the changing demands for energy in the world. Clearly this warrants considerable attention. The hope is that the information given here will encourage some readers to actually use these design concepts to make practical devices. A number of other semiconducting device designs that can be made using electroplating of silicon are reviewed next in the hope of further motivating the reader to consider electroplating as a general manufacturing option.

5.2.1 The Design of a Photovoltaic Structure

Maximizing the conversion of radiation that strikes a surface into electrical power implies that the photovoltaic material needs to have two features (1) good absorption of radiation and once a photon is absorbed (2) there should be a very short path for electron and hole to traverse to reach the current collectors. A short path is needed because impurities can stimulate recombination of the photoelectron and hole to generate heat instead of electrical power. The thinner the silicon, then the shorter is the path for charges to travel to the current collectors. The shorter the path, then the silicon can tolerate more impurities. However if the silicon is too thin, then radiation passes through it unabsorbed. A structure that promotes multiple reflections on the silicon surface can improve absorption in a thin silicon structure.

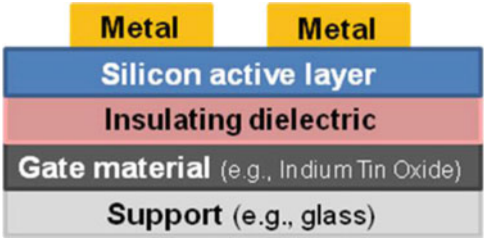


Fig. 5.2 Bottom gated thin film transistor

The photovoltaic device shown in Fig. 5.1 is the design of one structure which fulfills the two diametrically opposed needs for a thin yet efficient photovoltaic.

5.2.2 *Description of Thin Film Devices: TFT, NUC, and PV-NUC Hybrid*

Silicon electroplating offers an attractive alternative processing to conventional chemical processing of silicon. Electroplating gives a convenient way for forming thin films into complex geometries. A brief introduction to the variety of devices which can be made by electroplating is given next and then the silicon plating itself is discussed.

5.2.2.1 Thin Film Transistors

Perhaps one of the simplest devices that can be made by the electroplating of semiconducting silicon is a TFT. A TFT is a field-effect transistor, in which a voltage on the insulated gate electrode can induce a conducting channel between the two other metal contacts called source and drain. It is made by depositing thin film of a semiconductor active layer, and metallic contacts on an insulating dielectric layer and a gate material over a supporting substrate (see Fig. 5.2) [2]. The TFT is typically used for simple electronic circuits, such as video displays and power electronics.

5.2.2.2 Three-Dimensional Thin Film Structures

Figure 5.1 is a three-dimensional structure with a two-dimensional array of Schottky-diode photovoltaic diodes. These cylindrical diodes are cylinders (~2 μm in diameter) of metal and then silicon walls (~0.1 μm thick) filled with metal and all embedded in a thin (~60 μm thick) transparent membrane. A planar Schottky diode which has the same thickness of silicon as the silicon cylinder would transmit—not

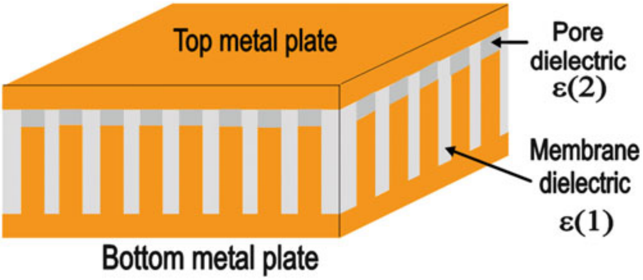


Fig. 5.3 Schematic diagram of a Nanowire Ultra Capacitor (NUC)

absorb—most of the impinging radiation. The three-dimensional structure allows multiple internal reflections on the silicon surface increasing the probability of photon capture leading to efficient captures of radiation. Another advantage of using a submicron cylinder of silicon is the probability of recombination of the photoelectron and hole is greatly reduced. The path to the metal current collectors is over a million times shorter so the purity of the silicon can be over a million times lower. Reducing the Si purity requirement gives the opportunity to use electroplating to make a photovoltaic. Electroplated silicon can have as much as 1 impurity atom per 10^6 silicon atoms, which is not nearly as pure as the 1 impurity atom per 10^{11} silicon atoms in the silicon made using CVD processes for the computer industry.

5.2.2.3 Nanowire Ultra Capacitor

A Nanowire Ultra Capacitor (NUC) is a three-dimensional thin-film device for electrical energy storage and for delivering electrical power and which is made by electroplating. A NUC consists of metal wires each of which are tens of microns in length and which are less than a micron in diameter embedded in a nanoporous dielectric membrane (see Fig. 5.3). The NUC behaves much like a dielectric capacitor, except that it has much higher energy density.

5.2.2.4 Renewable Hybrid Power Source (Photovoltaic and NUC Device)

A RHPS device is a parallel combination of an electrical-power-generating solar cell, like a three-dimensional photovoltaic (PV) Schottky diode shown in Fig. 5.1, and an electrical-energy-storing device like a NUC, shown in Fig. 5.3. This hybrid generator/storage device can be made by plating of silicon and copper in a nanoporous membrane. The structure of one RHPS repeat unit is schematically represented in Fig. 5.4.

Why use a PV-NUC hybrid design? There are two problems during PV power generation (1) one with line loss and (2) the other is intermittency. The line loss is a

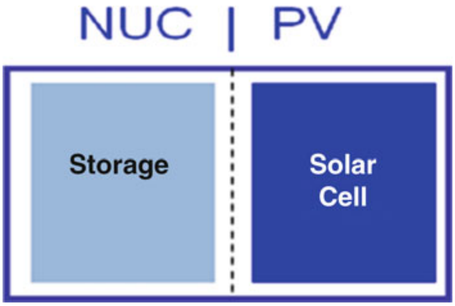


Fig. 5.4 Schematic diagram of a renewable hybrid power source (RHPS)

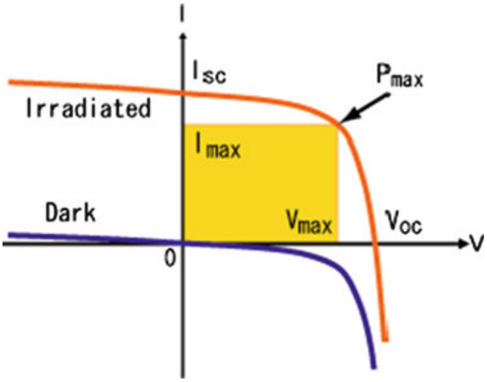


Fig. 5.5 The current–voltage characteristic of a PV device showing best conversion efficiency at $P_{\max}$ in the vicinity of the "knee"

more fundamental limitation. Line loss arises because the PV current–voltage characteristic is nonlinear (see Fig. 5.5). The PV exhibits a "knee" feature with diode-like behavior. The best photon to electrical energy conversion efficiency occurs when the operating point is in the vicinity of the "knee." PV power generation is inherently intermittent due to variation of ambient conditions like clouds and the sun rising and setting. A solution to both of these problems is the RHPS device, a hybrid of a PV device and storage device. The hybrid design maximizes solar power generation and avoids intermittency during solar power generation that could cause instability and other problems.

The distributed storage (here a NUC) minimizes resistance to PV conversion and allows maximal opportunity to hold the photovoltaic at $V_{\max}$, so power generation is at the knee voltage region. The presence of the NUC also allows steady power by discharge of NUC storage device during any intermittency in the PV power generation. A NUC, like any dielectric capacitor, can handle many thousands of charge/discharge cycles and is more reliable and durable than batteries for storing solar energy.

In summary, electroplating is a clean low-energy process that can be used to make photovoltaics and TFT by electrodeposition of silicon. Electroplating can be used to make an energy storing capacitor by plating metal. Accordingly, alternately electroplating in a monolith substrate can be used to make a hybrid device, a photovoltaic power generating device, and capacitive energy storing device, in one monolith. The key to doing all of this is to develop silicon electroplating, which is described next.

5.3 Silicon Electroplating

To date, semiconducting silicon is being made by the inefficient but effective "Czochralski Technique." This technique involves chemical vapor deposition (CVD) of silicon in which only about 10% of the silane (SiH_4) feed is pyrolyzed to silicon at between 750 and 1,000°C, followed by zone refining. This method is used because it yields high quality silicon, even though it is a chemically and energy inefficient "dirty" process.

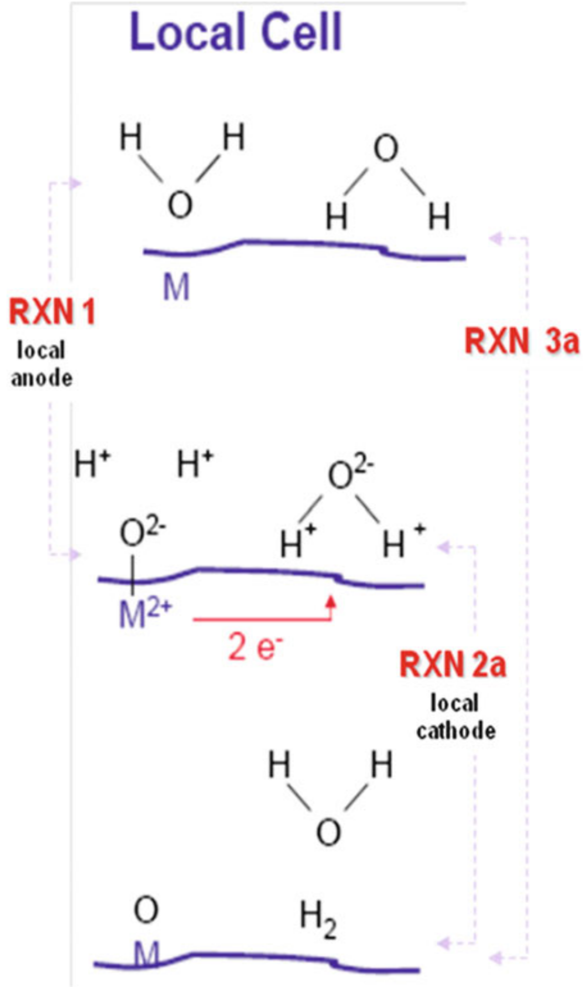
Relatively little practical work has been done in the area of electrodeposition of silicon [3]. This is almost certainly due to the intrinsic difficulty for electroreducing ionic silicon compounds to elemental silicon. The reduction potential for converting Si^{4+} to Si is more negative than the potential for the reduction of H^+ to H_2. Consequently in aqueous electrolytes, the evolution of H_2 is favored over Si formation. Even if silicon could be formed from the reduction of silicon salts in water, there is another problem. The newly formed silicon in water would immediately form silicon oxide. Metals, especially active metals like silicon, spontaneously react with water. The magnitude of the heat of formation (ΔH^f) of a metal oxide (see Table 5.1) is the driving force to form a metal oxide from a metal in the presence of water. The more negative the heat of formation, the greater the tendency to form metal oxide, as indicated in Table 5.1.

In aqueous electrolytes, a metal forms a metal oxide by the "Local Cell" process, which is schematically shown in Fig. 5.6. When water (H_2O) is in contact with the surface of a reactive metal (M), like aluminum or silicon, one part of the metal on the surface oxidizes (5.1) forming metal oxide (MO) while electrons travel through the bulk metal to another part of the metal surface where reduction reactions occur as shown in reactions 1 to 3 and the scheme in Fig. 5.6. Reactions 5.2a and 5.3a are for the anaerobic corrosion processes schematically shown in Fig. 5.6. Reactions 5.2b and 5.3b are for the analogous aerobic processes which are not shown.

Table 5.1 Heats of formation for conversion of metal plus water to metal oxide

	Metal oxide	ΔH^f (cal/g)
Reactive metals	Aluminum oxide (Al_2O_3)	−4,000 (= −1,676 kJ/mole)
	Silicon dioxide(SiO_2)	−3,418
Unreactive metals	Copper oxide (Cu_2O)	−278
	Platinum oxide (PtO_2)	−84 (= −80 kJ/mole)
	Gold oxide (Au_2O)	>0 (Au-oxide unstable)

Fig. 5.6 Metal oxide forming on a metal surface by the "Local Cell" process



$$M + H_2O \rightarrow MO + 2e^- + 2H^+ \quad (5.1)$$

The liberated electrons are consumed by proton to form hydrogen (5.2a). e^- can be consumed by oxygen to form water (5.2b).

$$2e^- + 2H^+ \rightarrow H_2 \quad (5.2a)$$

$$2e^- + \frac{1}{2}O_2 + H_2O \rightarrow 2OH^- \quad (5.2b)$$

Adding 5.1 with 5.2a (or 5.2b) gives the net undesirable metal oxide formation reaction, 5.3a (or 5.3b).

$$M + H_2O \rightarrow MO + H_2 + \text{heat} \tag{5.3a}$$

$$M + {}^1\!/\!{}_2O_2 \rightarrow MO + \text{heat} \tag{5.3b}$$

Metal oxide (MO) formation occurs with metal in the presence of water even in anaerobic environments (5.3a) because the electron (e^-) is more stable on H^+ than on the reactive metal (M) as shown in Fig. 5.6.

In the presence of oxygen from air, M is converted to MO (5.3b), because the e^- is more stable on O_2 than on M. Undesirable reactions (5.3a) and (5.3b) can be avoided in the absence of water. This can be achieved by using a nonprotic nonaqueous organic electrolyte or a nonprotic anhydrous ionic liquid electrolyte.

5.3.1 Silicon Plating from Nonaqueous Electrolytes

Water can be avoided by electroplating silicon from an aprotic organic-solvent or a molten-salt electrolyte [3]. In the 1980s, several researchers reported marginally successful electrodeposition of silicon in nonaqueous organic electrolyte baths, like trichlorosilane ($SiHCl_3$) in tetrahydrofuran (C_4H_8O) [4] and tetraethylorthosilicate (TEOS–$Si[OC_2H_5]_4$) in acetone [5]. In situ doping of Si during electrodeposition was described [6], using triethyl borate ($B(OC_2H_5)_3$) or triethyl phosphite ($P(OC_2H_5)_3$) for p-type or n-type material, although this doping was not particularly successful.

5.3.2 Silicon Plating from Organic Solvent Electrolytes

Many examples of electrodeposition of Si films from organic solvents were attempted and results are summarized. In 1981, Takeda et al. [5] used a plating bath of acetic acid as the solvent and tetraalkyl ammonium chloride added to aid the current flow and its distribution with tetraethylorthosilicate (TEOS–$Si(OC_2H_5)_4$) as the silicon source. Deposition was done on nickel substrates at room temperature, but the deposited films displayed infrared absorption spectra more typical of a silica gel rather than amorphous silicon or silicon dioxide. In 1981, Agrawal and Austin [7] used propylene carbonate and tetra alkyl ammonium chloride electrolyte with trichlorosilane ($SiHCl_3$) as the silicon source. Deposition was done at 35–145°C. The as-deposited silicon was amorphous, contained substantial amounts of hydrogen (16–35%) and oxygen (3%), and typically had resistivity exceeding 10^7 Ω cm. In 1982, Lee and Kroger [6] used a bath of acetone as the solvent, hydrofluoric acid (HF) as the supporting electrolyte, and potassium hexafluorosilicate (K_2SiF_6) as the silicon source. The as-grown films were reported to be amorphous Si, with resistivity exceeding 10^{11} Ω cm; the thickness of the films varied from 1 to 5 μm.

Experiments were also carried out to investigate in situ doping using triethyl borate ($B(OC_2H_5)_3$) or triethyl phosphite ($P(OC_2H_5)_3$) for p-type or n-type material, respectively. The doping efficacy was hampered by a compensating effect of the fluorine impurities in the electrodeposited films. In 1988, Gobet and Tannenberger [8] deposited amorphous silicon (a-Si) on metal or glass substrates by reducing trichlorosilane in tetrahydrofuran (C_4H_8O) with a supporting electrolyte. When the deposited Si film exceeded a thickness of 250 nm, it developed a network of cracks. Both oxygen and carbon were found in concentrations as high as 8%. The researchers were not able to determine whether the oxygen was incorporated during deposition or through exposure to air.

All of the above research groups mentioned that their investigation of electrodeposition of amorphous silicon was to establish a low temperature material growth process for low cost solar cells. However, none of these groups reported the successful fabrication of a solar cell. In all of the work listed earlier, the electrodeposited films are oxidized and almost certainly would not have been suitable for a junction solar cell.

Interest in electrodeposited Si has re-emerged recently, again. It is considered a possible alternative to amorphous Si grown by CVD, or gas discharge, or hot-wire deposition for low cost solar cells. Recognizing that oxidation of the electrodeposited film had to be eliminated, or at least substantially reduced, more recent efforts have employed different Si precursors, avoided the use of strong oxidizers in the electrolytes, and took stronger precautions to eliminate oxygen from the deposition environment:

In 2005, Nicholson [9] carried out Si electrodeposition from a bath of propylene carbonate solvent and tetra alkyl ammonium chloride (or bromide) as the supporting electrolyte, with either $SiCl_4$ or $SiBr_4$ as the silicon source. Small additions of PCl_3 or BCl_3 were used to incorporate n-type or p-type dopants, respectively. The deposited films still showed a strong tendency to oxidize. Secondary neutral mass spectrometry showed that the Si/O ratio in the deposition was approximately unity in most samples—except close to the interface between the deposited layer and the substrate. In 2005, Somberg [10] reported electrodeposition of a complete pn-junction structure on a Cu/Ni contact pattern electrolessly deposited on plastic (or ceramic) substrate. The p-layer and n-layer were deposited sequentially in separate electrolytic cells using propylene carbonate solvent, a tetraalkyl ammonium salt as the supporting electrolyte, a novel organic silicon fluoride salt as the silicon source, and unnamed dopant sources. The growth temperature was 40°C, and in situ sensors showed that the growth environment had oxygen and water concentrations less than 10 ppm. This particular report will be discussed in more detail later. In 2007, Nishimura and Fukunaka [11] carried out the Si electrodeposition with a bath of propylene carbonate solvent and tetrabutyl ammonium chloride as the supporting electrolyte with $SiCl_4$ as the silicon source. The deposited film was amorphous Si, and it was stated that upon exposure to air, the film oxidized rapidly, presumable due to its high porosity.

In spite of the precautions and modifications taken to reduce the oxidation of the deposited Si film, the results reported by Nicholson and by Nishimura and

Fukunaka are essentially the same and not altered from the results of the 1980s: the electrodeposited amorphous Si is amorphous and unstable and has substantial oxygen content. It is not clear whether the oxygen is introduced during the electrodeposition or upon exposure to air. Only the Somberg claim stands in sharp contrast to the results of both today's researchers and those from the 1980s.

5.3.3 Silicon Plating from Ionic Liquids

One other approach to electrodeposition, for both metals and semiconductors, employs a different type of nonaqueous solvent known as an ionic liquid (IL) [3]. Ionic liquids may be regarded as molten salts with a melting point below 100°C, and appear to fill the gap between the organic solvents, that tend to decompose above 100°C, and the traditional molten salts with melting points above 400°C. In 2004, Zein El Abedin and co-workers [12] described the electrodeposition of Si onto a gold-plated substrate in a room temperature ionic liquid (RTIL) electrolyte of 1-butyl-1-methylpyrrolidinium bis (trifluoro methyl sulfonyl) imide saturated with $SiCl_4$ solute. Since the ionic liquid is a good ionic conductor, no supporting electrolyte was added to the plating bath. The deposition was carried out at room temperature, and the electrodeposited silicon presented an interesting morphology, an open random clustering of small Si crystallites 50–100 nm in diameter stacked to a height of 500 nm. Additional work by this group demonstrated that the size and spatial distribution of the Si crystallites can be altered by adjusting the cell voltage and current [13–15]. Very little characterization of the electrodeposited silicon has been carried out. However, it did not appear that the Si displayed the tendency to oxidize as described with silicon plating in electrolytes with molecular organic solvents, above. No photovoltaic devices have yet been fabricated from Si deposited from this ionic liquid.

Recently, Gervasio et al. have developed a new approach to silicon electrodeposition [16] which is a variation of more recent silicon electroplating processes [12–14]. In short, a low viscosity ionic liquid electrolyte (*n*-butylpyridinium chloride) and a silicon source ($SiCl_4$, $SiHCl_3$, or Si_2Cl_6) are formed into a homogeneous mixture under an inert atmosphere at a temperature slightly above the boiling point of water. Plating is done at $T \geq 120°C$ because making certain that the silicon deposition occurs in a water free environment is a critical processing improvement over previous procedures. The exclusion of water prevents oxidation of the growing silicon films and should yield Si films suitable for high quality solar cells—although how effective this procedure remains is to be verified through the fabrication and testing of operational PV cells.

Silicon can be electroplated from a silicon salt continuously fed into an ionic liquid electrolyte or by "the electrowinning of silicon" process. The process of electrowinning of silicon is attractive because when "spiking" the electrolyte with silicon salts is avoided, so the species concentration should be more uniform in time for a more uniform deposition process. Electrowinning occurs when silicon salt is

$$\begin{array}{ll}
\text{IL rxn :} & 2\,R^{+}Cl^{-} + Si_2Cl_6 \rightarrow 2R^{+}SiCl_4^{-} \\
\text{Cathode rxn :} & R^{+}SiCl_4^{-} + 3\,e^{-} \rightarrow Si + R^{+}Cl^{-} + 3\,Cl^{-} \\
\text{Anode rxn :} & Si + R^{+}Cl^{-} + 3\,Cl^{-} \rightarrow R^{+}SiCl_4^{-} + 3\,e^{-}
\end{array}$$

Scheme 5.1 Electrowinning of silicon in ionic liquid electrolyte

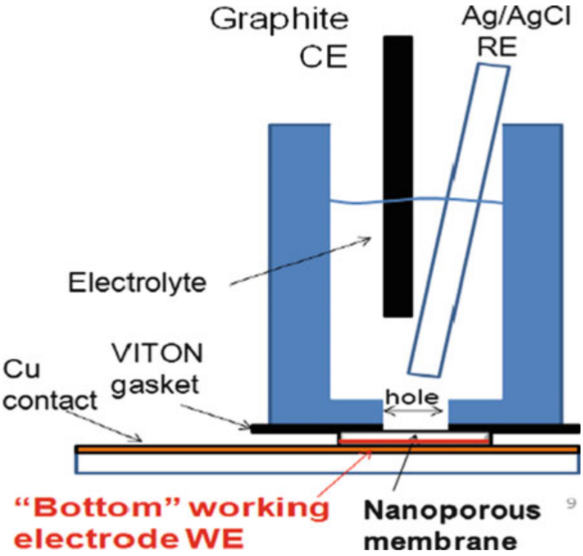
generated from a silicon anode held at a silicon oxidizing potential; the silicon then dissolves into the electrolyte and diffuses to the cathode. Silicon is deposited on a cathode held at a silicon reducing potential, as shown in Scheme 5.1.

In summary, the Si electrodeposition process consumes little energy, is environmentally friendly, and emits no volatile organic compounds (VOCs) or other noxious vapors over the reactor. The silicon-layer thickness is readily controlled by setting the plating current and time, and the deposition is typically conformal. This permits the construction of planar or textured device geometries. The electrodeposition process can be used to deposit single films or composite structures (silicon films with electroplated metal contacts and multiple silicon films to form a Schottky diode, pn-junctions or entire solar cell structures). Challenges include (a) establishing a repeatable and robust silicon eletrodeposition process, (b) attempting to dope the silicon layers with the judicious addition of boron and phosphorous precursors, and (c) attempting to grow multiple-layer structures with uniformity and benign interfaces. With development, the new approach to Si electrodeposition should produce films far superior to those obtained in earlier research efforts. Doping may be more difficult. If it is not possible to overcome the challenge of in situ doping, then other benign, low cost, low temperature silicon processing steps can be used, such as spinning or sputtering-on dopants or metal films, etc., to complete the processing of a silicon solar cell.

5.4 Preliminary Electrodeposition of Thin Films of Silicon for PVs

The electroplating of silicon from a nonaqueous electrolyte onto a surface is a low temperature (120°C) process which can be well controlled using the simple set up shown in Fig. 5.7. The setup consists of a cell made of glass or Teflon with three electrodes (1) the working electrode, which is the surface being Si plated; (2) the reference electrode, which is used to define or control the working electrode's potential using a voltmeter and a feedback loop; and (3) a counter electrode which passes oxidation current as the working electrode passes reducing current during the electroplating of silicon on the working electrode surface. The cell is filled with an electrolyte which introduces the silane reactant to the working electrode and completes the current loop between the working and counter

Fig. 5.7 Cell for electrodeposition of silicon on a conducting surface in 3 electrode configuration



electrodes inside the cell. An inert atmosphere can surround the cell to exclude undesirable species like oxygen and water. A heater can control the cell temperature to control the rate of reactions and to eliminate liquid water by heating above its boiling point since nonvolatile ionic liquids are used as the electrolytes.

Ionic liquids are becoming increasingly important materials with unique properties suitable for a wide variety of applications. The term "ionic liquid" refers to a molten salt that is liquid at temperatures equal to or above 100°C. These liquids are composed entirely of ionic species that is, ionic liquids consist of cations and anions alone and use no molecular species to form the electrolyte. In general, the cations are relatively complicated structures with their charge shielded and delocalized, which results in weak bonding and low melting points. The RTILs are liquid 25°C and above and are most useful for electrodeposition, RTILs employ large and asymmetric organic cations. The anions can be simple halides or more complicated structures. For example, the RTIL, 1-butyl-3-methylimidazolium hexafluorophosphate ([Bmim] PF_6) has the structure shown in Fig. 5.8.

Ionic liquids generally possess a number of desirable characteristics useful for silicon electrodeposition, such as high ionic conductivity, so no supporting electrolyte additions are needed, extremely low vapor pressure, which readily permits purification through evaporation of impurities, high thermal stability, large electrochemical potential window, typically around 5 V and excellent solvating properties for silicon solutes. Using ionic liquids, such as butyl pyridinium chloride and its derivatives, in place of organic solvents like THF, leads to greater stability in the electrolyte because ionic instead of molecular species are intrinsically more stable. Ionic liquids also dampen nonuniformity in the electric field on the plating surface which for all other things being equal is expected lead to a more uniform electrodeposited silicon layer on the surface. In general, electroplating of silicon

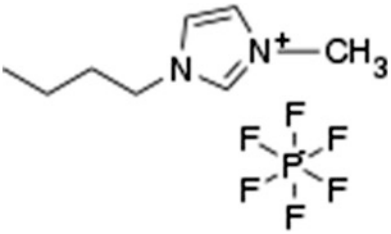


Fig. 5.8 Atomic structure of [Bmim] PF_6

is expected to lead to a polycrystalline silicon layer on an amorphous conducting substrate. Single crystalline silicon might be achieved by epitaxial electrodeposition on metalized mica (an inexpensive naturally occurring mineral which is well known to have a 111 crystallographic surface). This is highly speculative but would be interesting to put to the test.

5.4.1 Silicon Plating on Metal

Electrochemical measurements were carried out using a PARstat 2250 potentiostat–galvanostat. The experiments were conducted in a Teflon cell at room temperature in a glove box under a nitrogen atmosphere. The working electrode was a 0.127 mm × 0.5 cm × 3 cm titanium foil (Alfa Aesar Puratronic grade). The counter electrode consisted of a 0.5 mm in diameter platinum wire (Alfa Aesar Puratronic grade). The quasi-reference electrode was also a 0.5 mm diameter platinum wire. Potential in the experiments is reported as the potential difference from an internal reference, the ferrocene\ferrocenium redox couple. The ionic liquid, [Bmim]BF_4, was prepared by mixing equimolar quantities of 1-butyl-3 methylimidazolium chloride (Aldrich) and sodium tetrafluoroborate (Aldrich) in a methylene chloride solution [17]. The mixture was washed five times with deionized water to remove sodium chloride ions. The methylene chloride was removed via a rotary evaporator and the resulting ionic liquid was dried in a vacuum oven for 3 days at 120°C. The resulting liquid is a viscous, clear, pale yellow liquid. The titanium metal foil sample was characterized by scanning electron microscopy (SEM) and energy dispersive X-ray spectroscopy (EDAX).

5.4.2 Results and Discussion

Figure 5.9 shows the cyclic voltammogram for titanium foil in [Bmim]BF_4 ionic liquid with added ferrocene to establish the reference potential. There are no Faradaic processes between −2.5 and +1.5 V except for the oxidation of Ferrocene at around +0.4 V and reduction at around −0.4 V. The 0 of potential is the average of the oxidation and reduction potentials. The slight peak at 1.25 V may be

Fig. 5.9 Voltammogram of titanium foil in [Bmim] BF_4 with Fc but no $SiCl_4$ under N_2 atmosphere. Scan rate = 10 mV/s. T = 25°C

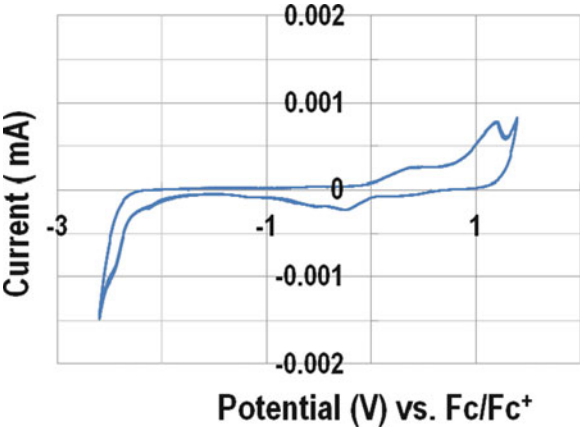
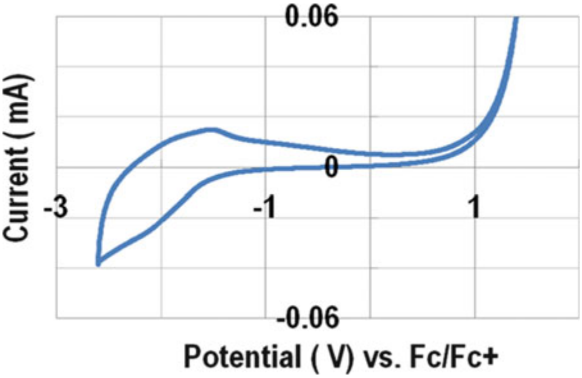


Fig. 5.10 Voltammogram of titanium in [Bmim]BF_4 saturated with $SiCl_4$ under 1 atm dry N_2. Scan rate = 10 mV/s. T = 25°C



oxidation of Pt to Pt oxide due to trace water. Above 1.5 and below −2.5 V, the ionic liquid oxidizes and reduces, respectively.

A fresh [Bmim]BF_4 liquid was then saturated with $SiCl_4$ and using the same setup then a cyclic voltammogram was measured and is shown in Fig. 5.10. Visual inspection of the differences in cathodic current in Figs. 5.9 and 5.10 indicates silicon plating was occurring at potential negative of −1 V versus the ferrocene reference.

Using the same cell, a constant potential of −2.0 V vs. Fc/Fc^+ was then applied to the working electrode for 10 min. The resulting chronoamperogram is shown in Fig. 5.11. The mass of the titanium working electrode was measured before and after the plating procedure, and it was found that there was a net gain in mass of 0.28 mg on the working electrode. Based on the current, plating time, and Faradays law of electrolysis, the expected mass of plated material based on the total charge passed was calculated to be 0.43 mg.

After the electrodeposition, the ionic liquid had turned a dark brown color, suggesting that something in this mixture of ionic liquid and $SiCl_4$ was changing during plating. Either (1) the ionic liquid is unstable to the potentials at the cathode

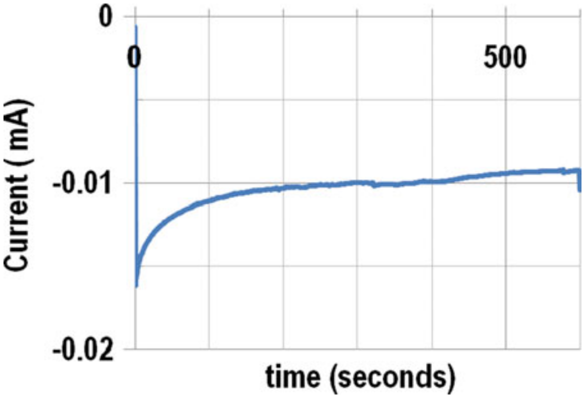


Fig. 5.11 Chronoamperometry of titanium metal in [Bmim]BF_4 saturated with $SiCl_4$ with *E* at −2.0 V vs. Fc/Fc^+ under a N_2 at 25°C

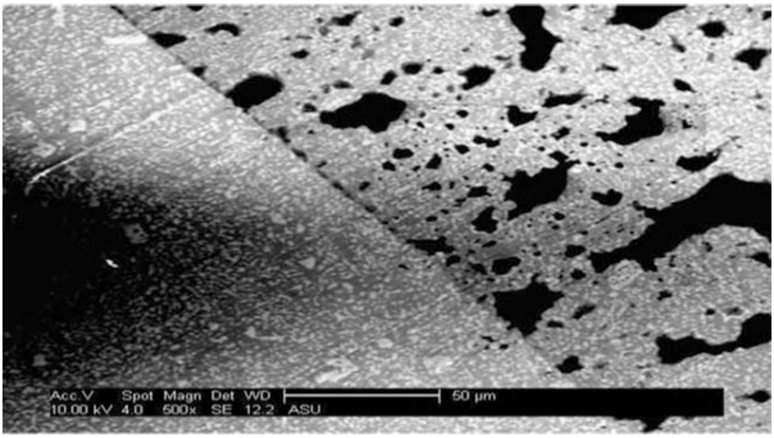


Fig. 5.12 SEM image of silicon on titanium and bare titanium substrate. *Bar* is 50 µm

or (2) chlorine is forming at the anode reacting with the ionic liquid or (3) the ionic liquid is not stable to anode potentials under these conditions. It appears the second possibility is the problem. Adding $SiCl_4$ introduces chloride to the electrolyte. The new large current which starts at ~1 V and is depolarizing the cell at 1.5 V versus ferrocene would appear chlorine generation and is most likely the source of color. This can be avoided by using nonchlorine silicon salts which is the subject of ongoing studies.

Figure 5.12 shows the SEM image of the bare titanium substrate (on the left) and the portion of the substrate where the titanium was submerged in the electrolyte and silicon was electrodeposited on the titanium substrate. The EDAX showed the coverage is over the whole surface on the right side but the SEM shows that the silicon plate is irregular in height. Leveling the plate height is the subject of ongoing studies.

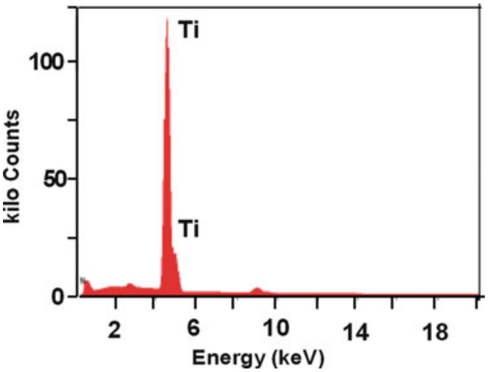


Fig. 5.13 EDAX spectrum for the bare Ti foil sample before Si electrodeposition

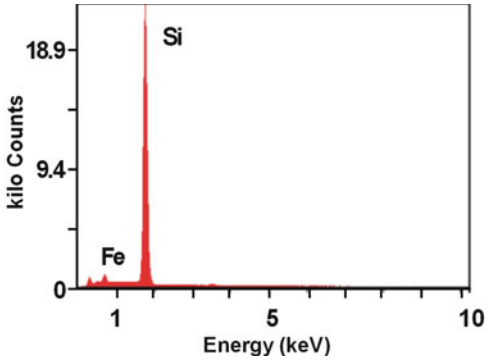


Fig. 5.14 EDAX spectrum for the Ti foil where Si was deposited at −2.0 V for 10 min

Figure 5.13 shows the EDAX spectrum for the titanium sample before it was used for silicon plating. There is only one peak in the EDAX spectrum shown in Fig. 5.13 at around the energy typical of Ti. This indicates that the Ti foil is at least 99% titanium.

Figure 5.14 contains the EDAX spectrum for the titanium sample after electrodeposition of silicon on titanium from the RTIL electrolyte, butyl, 3-methylimidazolium chloride, [Bmim] BF_4, which was in a glove box filled with dry nitrogen gas to exclude water. The EDAX spectrum of the portion of the titanium foil plated with silicon (see Fig. 5.14) shows virtually the presence of titanium only on the surface. This implies a complete coverage with a silicon film that is at least tens of nanometers in thickness. There is also a trace of iron which could have been introduced by the clips holding the cathode or anode.

5.5 Conclusions

The purpose of this chapter has been to review the state of solid state photovoltaics and suggest a strategy for lowering cost dramatically so that photovoltaics can become competitive with the cost of capital investment typically needed for supplying the grid with electrical power.

The suggested approach and recommendation is to use electrodeposition instead of CVD of silicon to make photovoltaics. Three main conclusions can be made about this approach:

1. Electroplating of silicon can be done from ionic silicon species dissolved in low temperature salts electrolytes called ionic liquids.
2. Ionic liquids are well suited for silicon electroplating because ILs are stable, do NOT need water or other reactive molecular solvents to be highly ion-conducting media, and can exclude solvents by plating at 120°C.
3. Electroplating of silicon from ionic liquid electrolytes provides a low-cost, energy-efficient, and clean alternative to CVD methods as a process for depositing high quality silicon and is a relatively simple way for forming silicon in complex structures needed for efficient photovoltaics.

For the reasons above, electroplating shows promise as a practical way to achieve suitable semiconducting silicon deposits at low enough cost to provide silicon photovoltaics, which are cost competitive with typical capital investment for large-scale grid power generation.

Acknowledgements The authors thank Jeffery Thomson, Telpriore Tucker, and John Gustafson for preliminary experiments on silicon electroplating and characterization, John Brews for discussion leading to design of photovoltaic structures based on silicon electroplating and the Office of Naval Research for support of portions of this work on capacitors under Project number H94003-09-2-0902, James Morris, Project Monitor and the Edson Venture Program of the State of Arizona for support of the work on electroplating of silicon.

References

1. http://www.solarbuzz.com/Moduleprices.htm; Lux Research Report “Balance of Systems: The Next Step to Grid Parity”. http://www.luxresearchinc.com
2. http://en.wikipedia.org/wiki/Thin-film_transistor
3. F. Endres, A.P. Abbott, D.R. MacFarlane, *Electrodeposition from Ionic Liquids* (Wiley-VCH, Weinheim, 2008)
4. J. Gobet, H. Tannenberger, J. Electrochem. Soc. **135**, 109 (1988)
5. Y. Takeda, R. Kanno, D. Yamamoto, T.R. Ram Mohan, C.H. Lee, F.A. Kroger, J. Electrochem. Soc. **128**, 1221 (1981)
6. C.H. Lee, F.A. Kroger, J. Electrochem. Soc. **129**, 936 (1982)
7. K. Agrawal, A.E. Austin, Electrodeposition of silicon from solutions of silicon halides in aprotic solvents. J. Electrochem. Soc. **128**(11), 2292 (1981)
8. J. Gobet, H. Tannenberger, Electrodeposition of silicon from a nonaqueous solvent. J. Electrochem. Soc. **135**(1), 109–112 (1988)
9. J.P. Nicholson, J. Electrochem. Soc. **152**(12), C795–C802 (2005)
10. H. Somberg, Production-scale electrodeposited silicon solar cells: issues to be solved for continuous manufacture. Paper presented at IEEE 4th world conference on photovoltaic energy conversion, Vol. 1, pp 1146–1147, 2006
11. Y. Nishimura, Y. Fukunaka, Electrochemical reduction of silicon chloride in a non-aqueous solvent. Electrochim. Acta **53**, 111–116 (2007)

12. S.Z. Abedin et al., Electrodeposition of metals and semiconductors in air- and water-stable ionic liquids. Chemphyschem **7**(1), 58–61 (2006)
13. S.Z. El Abedin et al., Electrodeposition of nanoscale silicon in a room temperature ionic liquid. Electrochem. Commun. **6**, 510–514 (2004)
14. S.Z. Abedin et al., Electrodeposition of metals and semiconductors in air- and water-stable ionic liquids. Chemphyschem **7**, 58 (2006)
15. N. Borisenko, S. Zein El Abedin, F. Endres, In situ STM investigation of gold reconstruction and of silicon electrodeposition on Au(111) in the room temperature ionic liquid 1-Butyl-1-methylpyrrolidinium Bis(trifluoromethylsulfonyl)imide. J. Phys. Chem. B **110**, 6250–6256 (2006)
16. D. Gervasio, J.Gustafson, Electroplating of semiconducting silicon. ASU Invention Disclosure, Case #: M8-042, 2007
17. P.A.Z. Suarez et al., Polyhedron Lett. **15**, 1217–1219 (1996)

Additional Reading

18. R. Abbashian, *Physical Metallurgy Principles*, 3rd edn. (PWS Publishing Company, Boston, 1994), pp. 62–66
19. L.G. Boxall, H.L. Jones, R.A. Osteryoung, Solvent equilibria of $AlCl_3$-NaCl melts. J. Electrochem. Soc. **120**(2), 223–231 (1973)
20. U. Cohen, Epitaxial growth of silicon or germanium by electrodeposition from molten salts. U.S. Patent 3983012, 1975
21. X. Deng, E.A. Schiff, Amorphous Silicon-Based Solar Cells, in *Handbook of Photovoltaic Science and Engineering*, ed. by A. Luque, S. Hegedus (Wiley, New York, 2003)
22. J. Gobet, H. Tannenberger, Electrodeposition of silicon from a nonaqueous solvent. J. Electrochem. Soc. **135**(1), 109–112 (1988)
23. R.A. Osteryoung, B.J. Welch, Electrochemical studies in low temperature molten salt systems containing aluminum chloride. J. Electrochem. Soc. **118**, 455–466 (1981)
24. R.D. Rauh, T.L. Rose, T.G. Hoover, D.L. Natwig, Electrodeposition of polycrystalline and amorphous silicon for photovoltaic applications, Quarterly Technical Progress Report No.1, November, 1979, DOE Division of Solar Energy, Contract DE-AC03-79ET23046
25. W.R. Runyan, in *Silicon and Silicon Alloys (Pure)*, vol. 24. Encyclopedia of Chemical Technology (Wiley, New York, 1997), pp. 1084–1093
26. B. Sopori, Thin-film Silicon Solar Cells, in *Handbook of Photovoltaic Science and Engineering*, ed. by A. Luque, S. Hegedus (Wiley, New York, 2003), p. 2003
27. J.T. Staley, W. Haupin, in *Aluminum and Alloys*, vol. 2, 4th edn. Encyclopedia of Chemical Technology (Wiley, New York, 1992), pp. 184–212
28. E. Yablonovitch, in *Electronic Materials*, vol. 9, 4th edn. Encyclopedia of Chemical Technology (Wiley, New York, 1997), p. 224
29. D.R. Lide, *CRC Handbook of Chemistry and Physics*, 85th edn. (CRC, New York, 2004), pp. 8–34

Chapter 6
Resistive Switching Models by Ion Migration in Metal Oxides

Daniele Ielmini

Abstract Resistive switching in metal oxides is considered one of the most promising storage concept for future generations of nanoscaled nonvolatile memories. In bipolar resistive switching, the resistance of a conductive filament (CF) is controlled through the application of electrical stimuli, and the conductive state of the nanoscaled CF can be used to encode the value of the logical bit in a nonvolatile memory. To investigate the scaling opportunities of the resistive switching concept, physical models must be developed. This chapter summarizes the current state understanding of bipolar resistive switching, providing evidence for the voltage across the device being the controlling parameter for the CF growth during set. A physical model for set and reset transition is then described, allowing for an interpretation of the observed switching characteristics for different timescales. Finally, the open challenges for the scaling and the reliability of resistive switching memories are briefly summarized.

6.1 Introduction

The resistive-switching memory (RRAM) is an emerging memory device based on the reversible change of resistance in an active material, usually a metal oxide (e.g., NiO, TiO_2, HfO_x, TaO_x) [1–5] or a chalcogenide glass (e.g., GeSe, GeS) [6, 7]. The change of resistance is generally due to the formation and the dissolution of a conductive filament (CF) with a size of a few nanometers. The CF is initially created by a dielectric breakdown process, similar to the one observed in silicon dioxide [8–10]. In metal oxides, the material degradation due to high field and high temperature during breakdown results in a local transformation of the insulating

D. Ielmini (✉)
Dipartimento di Elettronica e Informazione and IU.NET, Politecnico di Milano,
Piazza L. da Vinci 32, 20133 Milano (MI), Italy
e-mail: ielmini@elet.polimi.it

A. Korkin and D.J. Lockwood (eds.), *Nanoscale Applications for Information and Energy Systems*, Nanostructure Science and Technology,
DOI 10.1007/978-1-4614-5016-0_6, © Springer Science+Business Media New York 2013

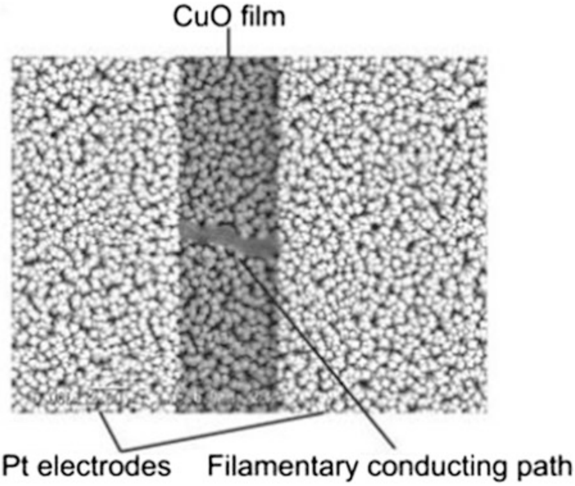


Fig. 6.1 SEM image of a conductive nanofilament, in a Pt–CuO–Pt resistive switching memories. Reprinted with permission from [2] (©2008 Elsevier Ltd.)

active layer into a conductive phase, e.g., a metal [11] or an oxygen-deficient region with relatively high conductivity [12, 13]. Figure 6.1 shows a scanning electron microscopy (SEM) image of a typical CF in a Pt–CuO–Pt RRAM, supporting the localized nature of the switching [2]. Other reports have indicated that the CF in TiO_2 RRAM consists of Ti_4O_7, a Magneli phase of titanium oxide exhibiting metallic conductivity [12]. For reference, oxygen-deficient silicon oxide was also found at the breakdown spot of the gate dielectric in MOS devices [9, 10]. In chalcogenide-based RRAM, instead, the CF originates from electrodeposition of cations, such as Ag and Cu, from one electrode to the other due to field- and temperature-induced electrochemical reactions [6, 7]. These electrochemical devices are generally referred to as conductive-bridge memory (CBRAM) and can also feature silicon or metal oxides as active layers or electrolytes [14, 15].

The first observations of reversible switching in metal oxides date back to the 1960s, demonstrating electrically induced resistance change in NiO [16], TiO_2 [17], and Nb_2O_5 [18]. Most recently, resistive switching was shown to take place in ternary oxides such as Cr-doped $SrTiO_3$ [19] and ferroelectric $Pb(Zr_{0.52}Ti_{0.48})O_3$ (PZT) [20]. In the last decade, intensive industrial and academic research on RRAM devices has resurfaced for several binary metal oxides, such as NiO [21–24], TiO_x [25–28], CuO_x [29], HfO_x [30, 31], and TaO_x [32, 33]. These works have aimed at evaluating the potential of the new technology in terms of scalability in view of a possible replacement of Flash memories for high-density nonvolatile storage and memories. From this standpoint, RRAM is most attractive for ultrascaled nonvolatile memories below the 10 nm node, due to its low and controllable programming current and fast switching, resulting in a low switching energy. For instance, a programming current below 10 μA was reported thanks to the control of the CF size through a transistor in series with the RRAM device, in the so-called one transistor/one resistor (1T1R) structure [34–37]. Programming currents in the range of a few μA have also been reported in other instances [38–40]. On the other hand, programming times in the range of a few ns or even less have been reported [40–42]. As a reference, considering a programming current of 1 μA with a programming time of 1 ns with a programming voltage of 1 V, a switching energy of 1 fJ is obtained. For comparison, the switching time in SRAM and DRAM is in the range of 3–10 ns with a switching energy of around 10 pJ, which highlights the strong potential of RRAM as a possible replacement of both nonvolatile and computer (system) memories. The area scaling of RRAM also seems promising, given the small size of the CF, together with the reported capability to control the CF size through limitation of the programming current during the formation stage [36, 37].

Although promising from several different aspects, including area scaling, switching speed, and energy consumption, RRAM still faces some fundamental limits. First, industrial development of RRAM will be possible only if this technology is demonstrated to be scalable enough to cover two or three technology nodes, otherwise the return of the industrial investments would not be sufficient. Second, reliability issues are still not completely understood, particularly in terms of cycling endurance, namely the number of program/erase cycles that can be achieved in one cell, and data retention, namely the maximum lifetime of the memory state. In particular, the low cycling endurance might critically limit the applicability of RRAM for system memories replacing SRAM and DRAM. To best address such scalability and reliability issues, the physical mechanisms must be thoroughly understood and physically based models must be developed, thus allowing for accurate prediction of scalability and for improvement and optimization of reliability.

This chapter will cover the current status about the understanding of switching and reliability mechanisms in bipolar switching RRAM. First, memory operation, including formation and program/erase switching in unipolar and bipolar memory devices, will be described. Then, the microscopic mechanisms for the resistive switching will be discussed based on experimental results obtained from HfO_x RRAM devices. A physical model for resistive switching will be shown, demonstrating the ability to predict switching parameters, such as resistance in the low resistance state and programming current as a function of operating conditions, such as the maximum current during CF formation. Finally, the reduction of switching time, switching energy, and CF size and their reliability implications will be addressed.

6.2 Unipolar, Bipolar, and Complementary Switching Characteristics

Figure 6.2 shows typical *I–V* characteristics including the resistive switching transitions for (a) unipolar, (b) bipolar, and (c) complementary switching operations. In all cases, two transitions can be identified, namely a set transition for achieving a low resistance state and a reset transition to achieve a high resistance state. The low and high resistance states will be referred to as set and reset states in the following.

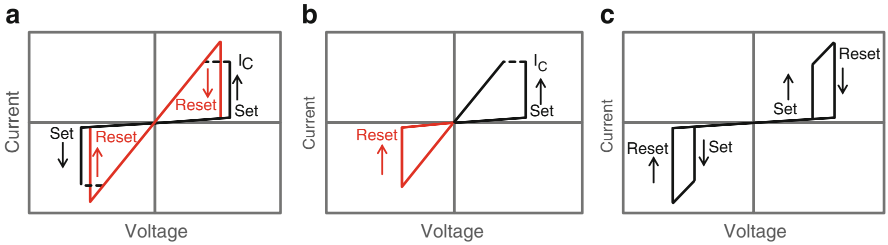


Fig. 6.2 Schematic current-voltage characteristics for the three switching modes of filamentary RRAMs, namely unipolar switching (**a**), bipolar switching (**b**), and complementary switching (**c**). Set/reset transitions occur independent of the bias polarity in unipolar switching, while they must be operated at opposite polarity in bipolar switching, e.g., set transition at positive polarity and reset transition at negative polarity. The maximum current flowing through the device after set is limited at a compliance current I_C in both unipolar and bipolar switching. In complementary switching, set with no I_C-limitation is followed by reset at the same polarity, allowing for the transition from a high resistive state to a different high resistive state. The two states differ by the position of the reservoir of conductive ions, see Fig. 6.3. Application of a negative sweep allows achieving the initial high resistive state through set and reset transitions

In unipolar switching (Fig. 6.2a), set and reset can be operated irrespective of the voltage polarity, while in bipolar switching set and reset are achieved under opposite polarities, e.g., positive voltage for set and negative voltage for reset (Fig. 6.2b). Complementary switching features instead set and reset transitions at both positive and negative polarity (Fig. 6.2c) [43]. Complementary switching was first introduced for ad hoc, antiserially connected RRAM stacks [44], then demonstrated as a fundamental switching mode in individual oxide layers in [43].

In the unipolar-switching case, the *I–V* curve of the reset state (OFF state in Fig. 6.2a) displays a sudden set transition at a relatively large voltage, where the device achieves a low resistance through formation (or re-activation) of the CF. In general, a current compliance I_C ("cc" in Fig. 6.2a), namely a limitation in the total current flowing through the device, is enforced during the set transition to prevent excessive growth of the CF which may eventually result in an irreversible breakdown [36, 37]. The *I–V* curve in the set state displays a reset transition at a reset voltage V_{reset} and a reset current I_{reset}. Reset is due to the dissolution of the CF, resulting in the transition to a high resistance state. No compliance is enforced during the reset current, thus allowing the necessary voltage and current to reach the values which are needed to dissolve the CF. The possible shape of the CF is schematically shown in Fig. 6.3 for the (a) set and (b) reset states. The conductive species are believed to diffuse radially or vertically from the CF at the hottest location during reset, as a result of thermally activated diffusion and migration [45].

In the bipolar switching case, the *I–V* curves in Fig. 6.2b display instead the set transition at a relatively large, positive voltage, while the reset transition takes place under a relatively low negative voltage. Note that the polarity may change depending on the particular active material/stack and on the initial forming

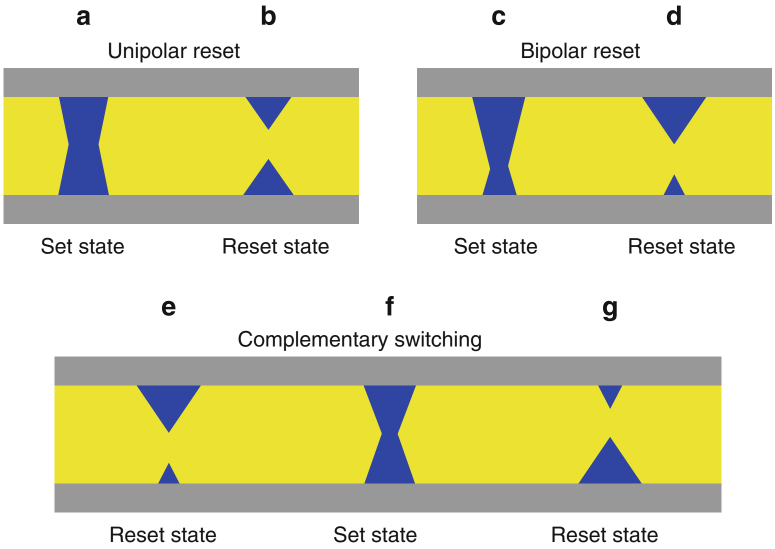


Fig. 6.3 Schematic illustration of the set and reset states in unipolar, bipolar, and complementary switching. Set state corresponds to a continuous CF in any case, while the reset state may differ depending on the polarity of the set/reset processes. In unipolar switching, set (**a**) and reset (**b**) states correspond to a continuous and interrupted CF, respectively, where filament breakdown is supposed to take place in the middle of the CF due to local temperature increase and consequent diffusion and oxidation. In bipolar switching, set (**c**) and reset (**d**) states correspond to a continuous and a interrupted filament, respectively, both with asymmetric structure due to the presence of a reservoir of conductive ions at one electrode (the top electrode in the scheme reported here). A negative applied voltage results in the transition from (**c**) to (**d**), as positively charged ions are attracted toward the top electrode by the electric field. In complementary switching, a positive applied voltage results in the transition from an initial reset state, corresponding to ions accumulated at the top electrode (**e**), to a set state, where ions are distributed along a continuous CF (**f**). A further increase of the applied positive voltage leads to the accumulation of conductive ions to the bottom electrode, thus resulting in a second reset state (**g**). Application of a voltage sweep with negative polarity results in the reverse transition, from (**g**) to (**e**) through set and reset transitions

operation. However, bipolar switching always requires voltage polarity reversal for repeatable set and reset processes. Figure 6.3c, d show the set and reset states, respectively, assuming that a positive (negative) voltage applied to the top electrode results in set (reset) transition. Most of the conductive ions are believed to remain at the top electrode, resulting in an asymmetric (e.g., conical) shape of the CF in the set state. Set and reset transitions mostly take place through vertical ion migration, activated by the temperature in the direction of the electric field.

Finally, the complementary switching characteristics in Fig. 6.2c were recently demonstrated for bipolar switching devices such as HfO_x RRAM [43]. In complementary switching, the set transition under positive voltage applied to the top electrode is not limited by any compliance, therefore the CF achieves a maximum size which is only limited by the amount of available conductive ions (e.g., hafnium ions Hf^+ or oxygen vacancies $V_O{}^+$) in the switching location. After completion of

CF growth, the voltage is further increased, leading to the migration of conductive ions toward the negative (bottom) electrode and to a consequent deactivation of the CF, resulting in a reset transition at a relatively large voltage. After reset, almost all ions that contributed to the formation and growth of the CF are located at the bottom negative electrode side and the resistance is relatively high, given the absence of a continuous CF. A similar operation is then carried out under negative voltage applied to the top electrode, leading to ion migration from the reservoir located at the bottom electrode toward the top electrode. This leads to a set transition, i.e., CF nucleation and growth, followed by reset, i.e., CF dissolution due to further ion migration toward the top negative electrode [43]. The same sequence can be further repeated until a maximum number of cycles, limited by the endurance lifetime. Complementary switching was first demonstrated in ad hoc stacks, consisting of a metal–insulator–metal–insulator–metal (MIMIM) structure where the intermediate electrode served as the initial reservoir for cations (positively charge ions), e.g., Cu or Ag ions [44, 46]. The application advantage of complementary switching is that the two states encoding the logic bit have both high resistances, since they consist of the conductive ions being accumulated either toward the top or the bottom electrode. The absence of a low resistance state in the array allows purely passive crossbar arrays to be achieved without the need for any select device, e.g., diodes or threshold switches [47–57]. The two states can be recognized through the sensing of the current response to a positive voltage at the top electrode: a set/reset sequence is observed if ions are located at the top electrode, but not if they are at the bottom electrode, thus allowing the recognition of the internal logic state. Note that reading is destructive, since ions are displaced from their original location. Figure 6.3e–g shows the CF shape in the high resistance state with conductive ions at the top electrode in the set state and in the high resistance state with conductive ions at the bottom electrode, respectively. The application of a positive voltage to the high resistance state in Fig. 6.3e results in ion migration toward the bottom electrode, leading to the formation of a continuous CF (set state in Fig. 6.3f) and to the accumulation of ions at the bottom electrode (reset state in Fig. 6.3g). The reverse path is followed if a negative sweep is applied to the state in Fig. 6.3g.

A general question regarding unipolar, bipolar, and complementary switching is what mechanisms drive the set/reset transitions and what is the connection between the existence of a certain type of switching and the active material properties. In general, unipolar switching is attributed to thermochemical mechanisms, namely physical mechanisms driven by the temperature alone, with little or no role of the electric field direction [4]. Set is explained as a chemical reduction of the metal oxide, e.g., NiO → Ni + O, due to the high temperature achieved at the CF site during set. Reset is instead generally explained as the inverse chemical reaction, namely oxidation, e.g., Ni + O → NiO, occurring at the metal-rich CF due to the large local temperatures during voltage application [58]. Note that oxidation and reduction may generally occur in the same material system (e.g., NiO) depending on the temperature range, i.e., reduction and oxidation requires a relatively high and low temperature, respectively [4, 59, 60]. For bipolar switching, instead, the driving force for switching is believed to be the ion migration induced by the electric field and accelerated by the local temperature due to Joule heating [45, 61]. Typical unipolar switching materials are NiO [4, 11, 21–24, 34, 36, 37, 49, 62, 63], TiO_2 [12, 27, 28], and TiON [64, 65], while bipolar switching materials include TiO_2 [25–27, 32], HfO_x [30, 31, 43], and WO_x [40]. In several cases, a coexistence between unipolar and bipolar switching has been observed, namely materials capable of bipolar switching were demonstrated to feature also unipolar switching [38, 66–69].

6.3 Bipolar Switching Characteristics

Figure 6.4 shows the measured *I–V* characteristics for a bipolar switching RRAM with HfO_x active oxide and TiN as electrode material in both the top and bottom contacts [70]. The thickness of the oxide layer was 20 nm, and the electrical switching was initiated by a forming stage at about +3 V (not shown). Set and reset transitions were observed under positive and negative voltage, respectively. The figure shows three *I–V* characteristics with a set/reset transition for increasing current compliance I_C enforced during the set transition, namely $I_C = 0.5, 0.75$, and 1 mA. The set transition occurs almost abruptly at about 0.5 V. Immediately after set, the current is limited to I_C while the voltage drop across the device reaches a value V_C of about 0.4 V. The resistance of the device in the set state is given by:

$$R = V_C/I_C, \tag{6.1}$$

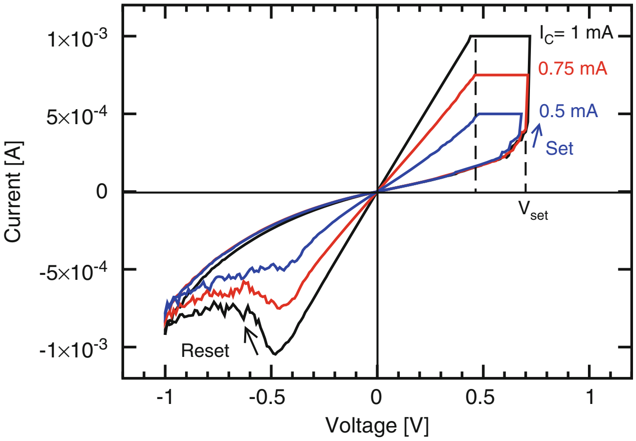


Fig. 6.4 Measured *I–V* characteristics for a bipolar switching RRAM, consisting of a TiN/HfO_x/TiN MIM stack with a 20 nm thick HfO_x layer. Set and reset take place under positive and negative polarity, respectively. Set takes place at V_{set}, then the current flowing during the set transition is externally limited to a compliance current I_C. Increasing I_C results in a decrease of the set-state resistance R and an increase of the reset current I_{reset}. Note the rather constant voltage $V_C = RI_C$ at the end of the set transition. The reset voltage V_{reset} is also almost constant and similar to V_C

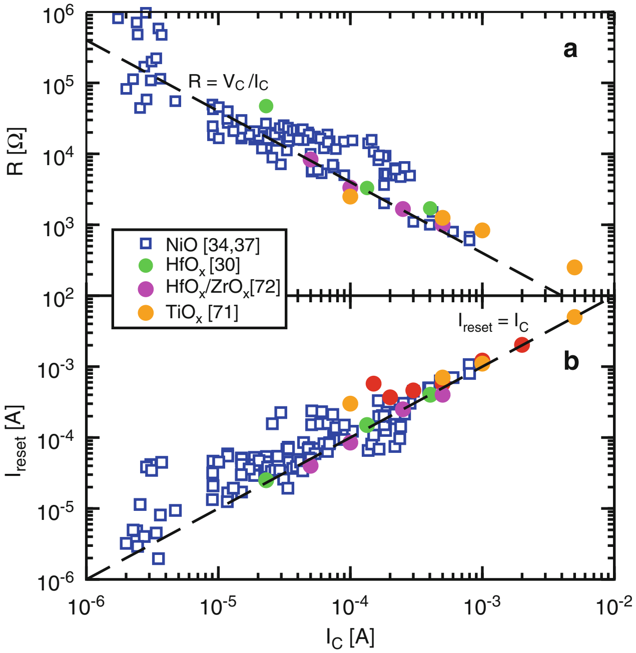


Fig. 6.5 Measured resistance R in the set state (**a**) and reset current I_{reset} (**b**) as a function of I_C, for both unipolar and bipolar switching and for several metal oxides, including NiO [21, 34, 36, 37], TiO_x [71], HfO_x [30], and HfO_x–ZrO_x [72]. The resistance is inversely proportional to I_C according to Eq. (6.1), while I_{reset} is proportional to I_C according to Eq. (6.2). Note that all data follow the same universal line, irrespective of the switching mode or active material. Reprinted with permission from [45] (© 2011 IEEE)

and is thus inversely proportional to I_C given the almost constant $V_C = 0.4$ V. At negative voltage, the reset transition is initiated at a reset voltage V_{reset}, which is almost constant around 0.4 V, thus roughly equal to V_C. The reset current I_{reset} is therefore approximately given by:

$$I_{\text{reset}} = \eta I_C, \tag{6.2}$$

with $\eta \approx 1$, given that $I_{\text{reset}} = V_{\text{reset}}/R \approx V_C/R = I_C$ [61].

These findings are generally observed in most oxide-based RRAM devices, as summarized in Fig. 6.5a, b [45, 61]. Figure 6.5a, b show the measured set state resistance and the reset current, respectively, as a function of I_C under DC-mode switching. Several oxide-RRAM materials are shown in the figure, including NiO [34, 37], TiO_x [71], HfO_x [30], and HfO_x–ZrO_x [72]. Data in Fig. 6.5a agree with the inverse proportionality between R and I_C in Eq. (6.1), with $V_C = 0.4$ V, while data in Fig. 6.5b confirm Eq. (6.2). The figures also report data for one transistor/one resistor (1T1R) structures, where the RRAM was in series with a MOS transistor to control the current flowing in the device during set [30, 34, 36, 37]. Data from 1T1R devices generally feature a smaller I_{reset}, thanks to the absence of current overshoot exceeding I_{C} during the set transition [36, 61, 73]. Note that data in the figures align on the same trend irrespective of the active material (NiO, HfO_x, TiO_x, etc.), of the switching mode (unipolar or bipolar switching), and of the cell structure (single MIM, 1T1R device). The independence of V_{C} and η in Eqs. (6.1) and (6.2) on the active material suggests a *universal* switching kinetic for the formation of the CF, with little or no dependence on the metal oxide composition [45, 61].

6.4 Time-Resolved Switching Characteristics

To gain more insight into the set transition described by Eq. (6.1) and into the universal switching characteristic, where V_{C} does not significantly depend on the material or switching mode, time-resolved switching experiments have been carried out. In these experiments, set pulses with voltage amplitude V_{A} were applied to a RRAM device with a load resistance R_{L} in series, as illustrated in Fig. 6.6. The initial state was a reset state with resistance around 5 kΩ. A sequence of pulses with increasing pulsewidth was applied and the resistance R of the device was measured after each pulse. Between each set pulse, no reset pulse was applied, thus allowing tracking of the cumulative effect of set pulses for increasing time. The voltage across the cell during each set pulse V was also evaluated as:

$$V = \frac{R}{R + R_{\mathrm{L}}} V_{\mathrm{A}}, \tag{6.3}$$

where R is the resistance measured after the corresponding set pulse.

Figure 6.7 shows the measured R as a function of time for $R_{\mathrm{L}} = 1$ kΩ (a), 2.2 kΩ (b) and 5 kΩ (c) and for increasing applied voltage V_{A}, from 1 to 3.5 V, for HfO_x-based RRAM devices [74]. Figure 6.7d–f show the corresponding estimated V

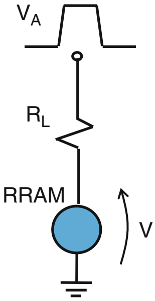


Fig. 6.6 Schematic illustration of the device layout for the time-resolved measurement of the set transition. The RRAM device is connected with a load resistance R_{L} to limit the current during set. A pulse with voltage V_{A} is applied, with a voltage drop V at the RRAM device. Reprinted with permission from [74] (©2011 IEEE)

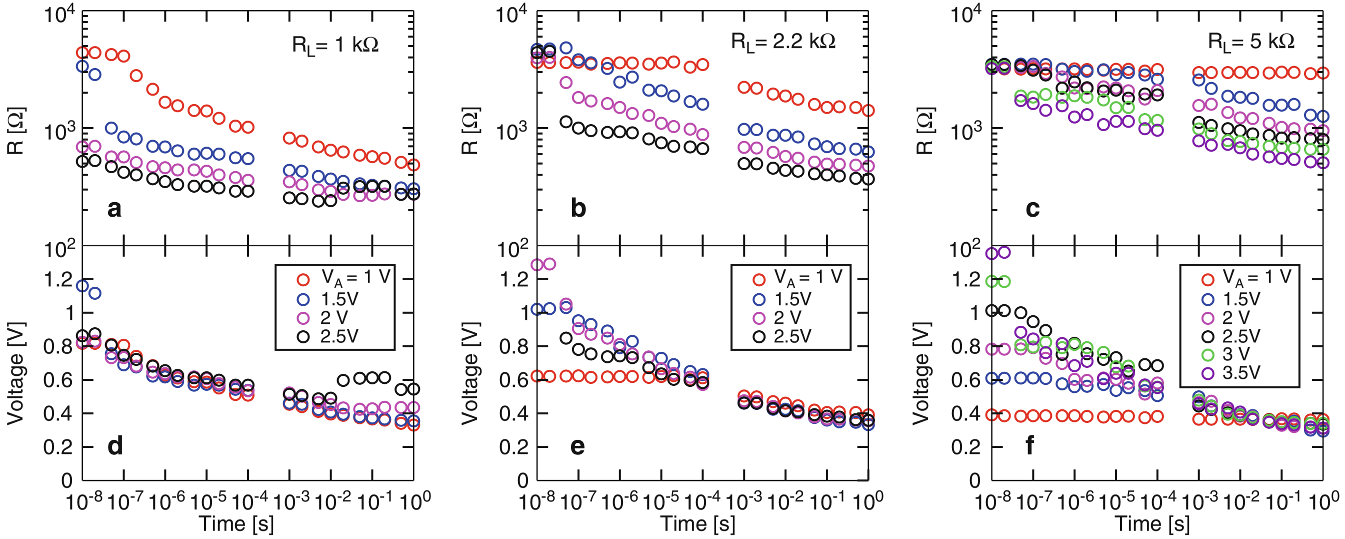


Fig. 6.7 Measured R and V for increasing load resistance $R_L = 1$ kΩ (**a**, **d**), 2.2 kΩ (**b**, **e**), and 5 kΩ (**c**, **f**). The resistance was measured in a broad time scale from 10 ns to 1 s at increasing V_A. The resistance drop reflects the formation and growth of the CF, where the growth rate increases for increasing V_A. The final R increases with R_L due to the decreasing compliance current. Note the universal character of the measured V across the cell, suggesting that V is the controlling parameter for the CF growth process. Reprinted with permission from [74] (© 2011 IEEE)

across the RRAM. Note the extremely wide time range in the experiments, from 100 ns to 1 s, which was possible thanks to a logarithmic increase of pulsewidth in the sequential set experiment [74]. In general, R decreases for increasing set time, as a result of the formation (nucleation) and growth of the CF. The set process becomes faster for increasing V_A, as expected due to the voltage-accelerated kinetics of filament growth. The resistance at the end of the set transient at 1 s increases for increasing R_L: This is similar to the compliance effect seen in Fig. 6.4, where a decreasing I_C (corresponding to an increase of R_L, since a larger R_L provide more limitation to the current through the device) results in a larger R at the end of set.

The resistance limitation effect in Fig. 6.7a–c can be understood by considering the time and V_A dependence of the voltage V across the cell as a function of time in Fig. 6.7d–f. Here, the RRAM voltage follows a unique function of time, which is independent from V_A and R_L. These results reveal that voltage is the controlling parameter for the filament growth process during set. In fact, V adjusts to a given value $V_C(t)$ at any time, irrespective of V_A and R_L, where V_C is dictated by the filament growth kinetics in the material. The self-adjustment of the RRAM voltage is due to the presence of a negative feedback by the presence of a load resistance in series with the cell [75]: as the CF grows, its resistance decreases, thus the voltage across the device decreases due to Eq. (6.3). The growth process thus quenches itself, resulting in a self-limiting growth kinetics. Note that, if no load resistance is put in series with the RRAM device, no negative feedback is enforced, thus leading, in principle, to an unlimited growth of the CF. (In real devices, other limitation mechanisms take place, as already demonstrated by the complementary switching behavior described in Sect. 6.3.)

The results in Fig. 6.7d–f highlight the meaning of the constant V_C at the end of the set transient in DC set experiments in Figs. 6.4 and 6.5: In particular, note in Fig. 6.7d–f that V is close to $V_C = 0.4$ V at $t = 1$ s, at the end of the set experiment. V_C has therefore the meaning of the natural voltage across the cell, which results from the CF growth in presence of a negative feedback, due, e.g., to a load resistance R_L in series with the cell in Fig. 6.7, or to a current compliance I_C in Figs. 6.4 and 6.5. Data in Fig. 6.7d–f also reveal that V_C depends on the set time, e.g., a larger V_C is expected for decreasing set time.

The universal behavior of V in Fig. 6.7d–f provides evidence for a voltage-controlled growth process in oxide-based RRAM. The voltage-controlled nature of the set transition results from the strong (e.g., exponential) dependence of the growth rate on voltage. In fact, such a strong dependence allows for a quick readjustment of the voltage across the cell at any given time during set. To gain more insight into the fundamental law governing the voltage dependence of the CF growth process, we have measured the set time t_{set} and the reset time t_{reset} as a function of cell voltage V. Set/reset times were estimated as the times for which the resistance decreases/increases by 50 % from the initial value. Figure 6.8 shows the time-resolved resistance change during reset in HfO_x RRAM, starting from a low-resistance set state and for increasing V_A between 0.6 and 1.2 V. No load resistance was applied during reset as already noted in Sect. 6.2, so that $V = V_A$. The reset time decreases for increasing V_A, revealing a similar voltage-controlled nature as evidenced for the set transition. Figure 6.9a shows the measured t_{set}/t_{reset} as a

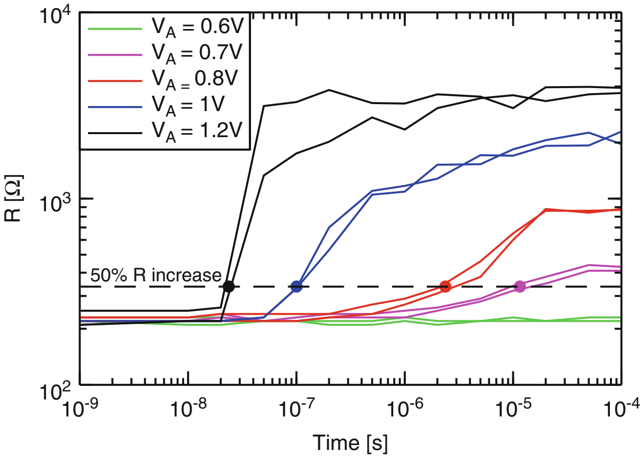


Fig. 6.8 Measured R as a function of time during reset for increasing V_A. No load resistance was applied in series with the cell. The reset time, corresponding to a 50% increase of R, is marked on the figure, indicating an almost exponential decrease of t_{reset} for increasing voltage

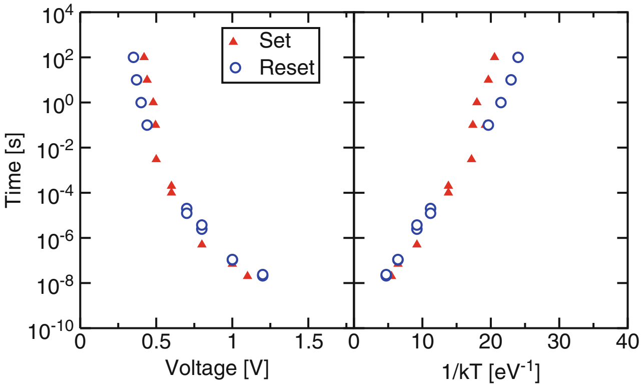


Fig. 6.9 Measured set and reset times, evaluated as a 50% resistance decrease or increase, respectively, as a function of voltage (**a**) and as a function of $1/kT$ (**b**), where T is the local temperature at the CF evaluated from the Joule heating formula of Eq. (6.4). The exponential dependence on $1/kT$ suggests that set/reset processes are due to temperature-activated mechanisms obeying an Arrhenius law

function of the cell voltage V, estimated by Eq. (6.3) with $R_L = 0$ for reset. Both t_{set} and t_{reset} follow approximately the same dependence, with a strong dependence on voltage. In fact, t_{set} and t_{reset} change by ten decades upon a voltage increase from about 0.3 V to about 1.25 V, e.g., roughly a factor 4 [74]. These results are in agreement with similar data for RRAM devices based on the migration of Ag cations [75] and based on TiO_x resistive switching [3]. Note, in particular, that the t_{set} and t_{reset} does not feature an exponential behavior on voltage, thus indicating that a simple F-model for set/reset is not sufficiently accurate. To understand the fundamental law driving set and reset kinetics, t_{set} and t_{reset} are shown in Fig. 6.9b as a function of $1/kT$, where k is the Boltzmann constant and T is the local temperature at the localized filament [74]. The local temperature was evaluated from the Joule heating model described by:

$$T = T_0 + \frac{R_{\mathrm{th}}}{R} V^2, \tag{6.4}$$

where R_{th} is the effective thermal resistance at the localized CF and T_0 is the temperature of the device (i.e., room temperature) [62, 75, 76]. The thermal resistance R_{th} was evaluated as:

$$R_{\mathrm{th}} = \frac{1}{8k_{\mathrm{th}}} \frac{4L}{\pi\phi^2}, \tag{6.5}$$

where k_{th} is the thermal conductivity in the CF, L is its length, assumed equal to the oxide thickness, and ϕ is the diameter of the filament, assumed to have a cylindrical shape. The factor 8 in the denominator of Eq. (6.5) is obtained by solving the Fourier differential equation in the presence of a distributed Joule dissipation [76, 77]. The electrical resistance is given by a formula similar to Eq. (6.5), except for replacing $1/(8k_{\mathrm{th}})$ with the electrical resistivity ρ. As a result, the ratio between thermal and electrical resistance in Eq. (6.4) can be written as:

$$\frac{R_{\mathrm{th}}}{R} = \frac{1}{8\rho k_{\mathrm{th}}}. \tag{6.6}$$

For the T estimation in Fig. 6.9, a ratio $R_{\mathrm{th}}/R = 1{,}500\ \mathrm{kV}^{-2}$ was assumed, corresponding to a thermal conductivity $k_{\mathrm{th}} = 20\ \mathrm{JK}^{-1}\ \mathrm{cm}^{-1}\ \mathrm{s}^{-1}$, similar to the thermal conductivity of bulk hafnium, and an electrical resistivity $\rho = 400\ \mu\Omega\ \mathrm{cm}$, which is about 30 times larger than bulk hafnium value. Such a large enhancement of electrical resistivity can be understood by the nonmetallic character of the CF, generally associated with a suboxide phase such a $\mathrm{Ti_4O_7}$ in the case of TiO_x RRAM [12, 13]. Also, one should consider that the electrical resistivity may be largely affected by significant surface and defect scattering at the nanoscale CF, as a result of the Fuchs–Sondheimer formula [78]:

$$\rho = \rho_{\mathrm{bulk}} \left(1 + \frac{3}{4} \frac{\lambda}{\phi} (1 - p) \right), \tag{6.7}$$

where ρ_{bulk} is the bulk electrical resistivity, λ is the electron mean free path, and p is the specularity factor, namely, the probability for elastic scattering at the surface of the CF. For instance, an electrical resistivity $\rho = 380\ \mu\Omega\ \mathrm{cm}$, and thus close to the value used to estimate R_{th}/R in Eq. (6.6), can be obtained for a Hf nanoscale filament of size $\phi = 1$ nm, assuming $\lambda = 28$ nm and $p = 0.5$ [61, 78]. This supports the choice of electrical resistivity in the CF.

Data for t_{set} and t_{reset} in the Arrhenius plot of Fig. 6.9b display an almost linear behavior, thus suggesting that the set/reset process for filament growth and dissolution obeys a strong temperature activation, driven by the local temperature increase due to Joule heating. Also, the similarity between t_{set} and t_{reset} suggests that set/reset mechanisms have the same nature, e.g., the temperature-activated ion migration driven by the electric field at a localized spot. The voltage application during the set/reset pulse might thus have two important roles (1) providing the necessary Joule heating to allow completion of the filament growth/dissolution in a sufficiently short time, thanks to the Arrhenius temperature activation in Fig. 6.9b, and (2) providing the necessary electric field needed to drive ion migration in the correct direction, thus allowing to either fill an existing high-resistance gap through ion migration, to nucleate and grow a CF during set, or open a high resistance gap by ion migration during reset. Note finally that set and reset times display a weak difference at small voltages and temperatures. In particular, t_{set} is generally larger than t_{reset} at small voltages, or, equivalently, the set voltage V_{set} is generally larger than the reset voltage V_{reset} for the same time scale of the set/reset experiment. This is also clear from Fig. 6.4, where V_{set} is around 0.6 V, with a smaller V_{reset} of about 0.44 V. This can be understood by the fact that set is initiated by nucleation, where the CF must be initially formed through ion migration across a high resistance gap. Such an initial condition differs significantly from the one in the reset process, starting from a continuous CF with low resistance. In particular, one may expect that the initial temperature is larger in the continuous CF with respect to the disconnected CF, as a result of the change in electrical resistance in Eq. (6.4): the relatively larger value of R in the set state, compared to the reset state, results in a smaller local temperature due to Joule heating, for the same applied voltage. In addition, one should consider that, in the set state, only the ion migration is accelerated by the temperature at the tip of the broken CF, which may be significantly smaller than the maximum temperature evaluated by Eq. (6.4). As a result, a resistance-dependent overvoltage must be supplied to the reset state to initiate the set transition. Generally, the larger the resistance, the larger the voltage needed to trigger set, according to Eq. (6.4). Note that one should also consider the change in the thermal resistance, which might possibly compensate the increase of R in Eq. (6.4). However, thermal resistance changes are believed to be less marked, due to the phonon contribution to thermal conduction in the quasi-insulating gap.

6.5 Physical Mechanism for Bipolar Switching

The previous experimental analyses have pointed out the following evidences:

- Set is a voltage-controlled process, where the filament growth kinetics is a strong function of voltage, thus the voltage across the cell follows a universal evolution with time in the presence of a load resistance or current compliance system.

- Set and reset processes share the same fundamental physical mechanism, since similar set/reset times are observed as a function of voltage across the device.
- Set and reset processes are temperature accelerated through an Arrhenius law, where the local temperature increase due to Joule heating allows for the completion of CF growth/dissolution mechanisms in an extremely short time, e.g., few ns or even sub-ns time scales.

These evidences point to a fundamental role of ion migration during set/reset processes. Ion migration, in fact, accounts for the bipolar character of switching, since ions are pushed and retracted toward one electrode during set and reset process, respectively, thus allowing for the filling and opening of a depleted gap with high resistance. Ion migration obeys also a temperature-accelerated kinetics, due to thermally activated ion hopping among localized states [61, 75, 79, 80]. In this perspective, it should be noted that Eqs. (6.1) and (6.2), which control the I_C dependence of set and reset voltage and currents, are found not only in oxide-based RRAMs (e.g., see Fig. 6.5), but also in conductive-bridge RAM (CBRAM), namely RRAM devices which rely on the migration of Ag or Cu. Such metallic ions are supplied from one of the electrodes and migrate through a high resistance layer, typically a chalcogenide glass [6, 7, 75] or an oxide layer [14, 15]. For reference, Fig. 6.10 shows the measured R (a) and the measured I_{reset} (b) as a function of I_C, for CBRAM devices based on Ag migration in GeSe and on Cu migration in SiO_x [15]. Data for HfO_x RRAM are also shown for reference [30]. Results obey Eqs. (6.1) and (6.2), although with $V_C \approx 0.2$ V and $\eta \approx 0.5$, and thus smaller than the values obtained from oxide-RRAM in Fig. 6.5. While the smaller values of V_C and η can be understood by different values of the activation energy for set and reset, the validity of Eqs. (6.1) and (6.2) suggests that the same physical mechanisms control set/reset processes in both oxide-based RRAM and CBRAM. Since the latter obviously rely on ion migration for CF formation/disruption, the same mechanisms is strongly supported as the key physical phenomenon driving set/reset in oxide-RRAM.

Figure 6.11 shows the physical picture for resistive switching in oxide-based RRAM, which is compatible with the experimental evidence gained so far. The figure shows snapshots during the set process, illustrating the nucleation (a, b) and growth (c, d) stages of the CF. The initial state corresponds to a broken CF, where conductive species (e.g., Hf ions and/or oxygen ions and/or oxygen vacancies in HfO_x) have been accumulated toward the top electrode by a previous reset process under negative voltage. The application of a positive voltage at the top electrode drives the ions toward the bottom electrode, resulting in the formation of a CF nucleus, i.e., the smallest continuous connection of the two electrodes through conductive species, and its subsequent growth. Electrical conduction through the CF results in Joule heating and a temperature increase, thus accelerating the growth process. As the filament grows, its resistance decreases, thus resulting in a decrease of the voltage across the oxide layer according to Eq. (6.3), if a compliance system or a load resistance is connected to the device during set. The resulting voltage reduction quenches the driving force for ion migration, thus slowing down the growth process. If no compliance or series resistance is present in the circuit, the

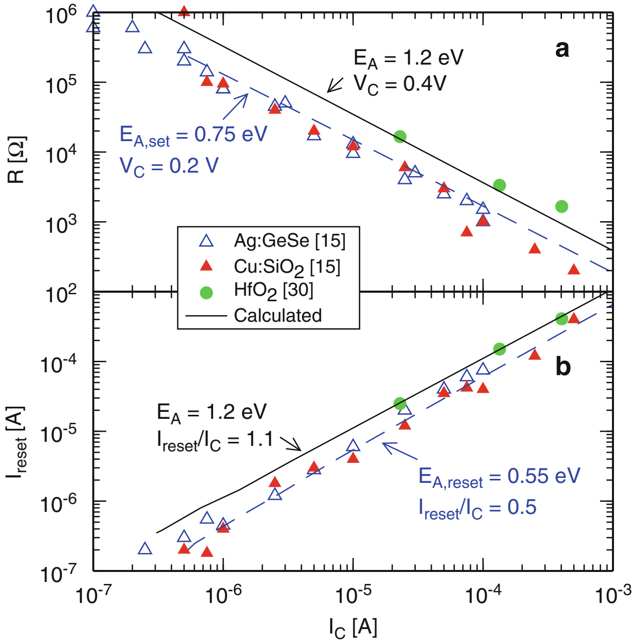


Fig. 6.10 Measured and calculated R (**a**) and I_{reset} (**b**) as a function of I_C, for HfO$_x$-based RRAM devices [30] and CBRAM devices based on Ag:GeSe or Cu:SiO$_2$ [15]. All material systems display similar behaviors in agreement with Eqs. (6.1) and (6.2), suggesting common mechanisms for set/reset in RRAM and CBRAM devices. The figure also shows calculations according to the model described in Sect. 6.6. Reprinted with permission from [74] (© 2011 IEEE)

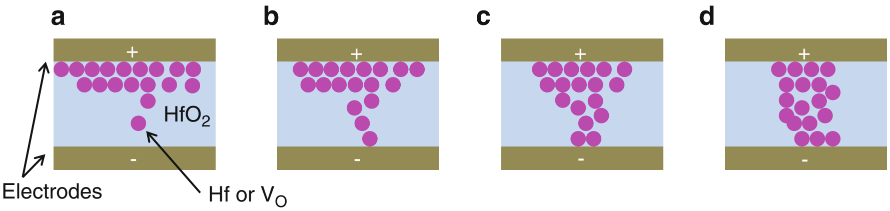


Fig. 6.11 Schematic illustration of the filamentary growth process based on ion migration during set. In a reference HfO$_2$ RRAM, ion migration results in filament nucleation (**a**, **b**) and successive growth (**c**, **d**), leading to a resistance drop during set transition. The reverse sequence applies for describing the reset process of filament dissolution. Reprinted with permission from [61] (© 2011 IEEE)

growth process will take place until all ions in the reservoir at the top electrode are used in the CF. From this point, further ion migration might result in a reduction of the CF size and in accumulation of the ions toward the bottom electrode, thus resulting in a reset process. This has been indeed demonstrated in HfO$_x$-based RRAM devices and is the basis for the complementary switching operation described in Sect. 6.1. For an interrupted set process described by Fig. 6.11, the reset process can be described by the same sequence but in reverse order, namely from d to a.

6.6 Physical Model for Bipolar Switching

The set transition illustrated in Fig. 6.11 can be given a simple analytical description through the experimental evidence for ion migration driven by field and the local temperature [61, 74]. According to this model, the size of the CF increases with time according to the following growth equation:

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = A\mathrm{e}^{-\frac{E_{\mathrm{A}}}{kT}}, \tag{6.8}$$

where E_{A} is the activation energy for ion migration and A is a preexponential coefficient. The model is justified by noting that the filament growth takes place through ion migration from the reservoir to the CF, and thus is limited by the migration process. The latter is thermally activated through the Arrhenius law, as already evidenced from experiments shown in Fig. 6.9. From this perspective, A may be a rather complicated function of ion diffusivity in the considered active oxide or grain boundaries, the filament length L, and the composition/volume of the ion reservoir (pure metallic phase, oxygen deficient oxide, etc.). The local temperature T in Eq. (6.8) can be obtained by the Joule heating in Eq. (6.4), while the activation energy decreases due to the applied field as a result of the barrier lowering effect. This is shown in Fig. 6.12, where the potential profile along the

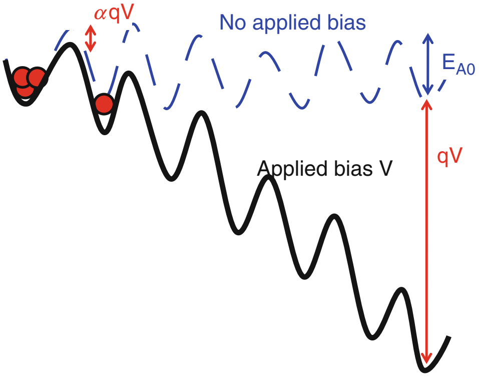


Fig. 6.12 Schematic illustration of the voltage-induced lowering of the energy barrier for ion hopping during filamentary growth/dissolution. Ions hop among potential wells through thermal excitation over potential barriers of amplitude E_{A0}. The voltage V lowers the barrier by an amount $q\alpha V$, thus enhancing migration in the direction of the electric field. Reprinted with permission from [45] (© 2011 IEEE)

active oxide layer is depicted for zero and nonzero applied voltages [45, 75]. Application of a voltage V across the oxide layer results in a lowering of the barrier according to:

$$E_{\mathrm{A}} = E_{\mathrm{A0}} - q\alpha V, \tag{6.9}$$

where E_{A0} is the zero-field activation energy and α is a constant [45, 75]. Substitution of Eqs. (6.4) and (6.9) in Eq. (6.8) results in:

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = A\mathrm{e}^{-\frac{E_{\mathrm{A0}}-q\alpha V}{k\left(T_0+R_{\mathrm{th}}/RV^2\right)}}, \tag{6.10}$$

which shows that application of a voltage has two enhancement effects on the growth rate, namely the barrier lowering and the Joule heating. Note that, while the model fully describes the growth process, it cannot account for the nucleation process, where the initial high-resistance gap is filled by ions. However, this approximation is largely acceptable because (1) the nucleation process only results in a minor overvoltage of V_{set} with respect to V_{reset}, of about $V_{\mathrm{set}} - V_{\mathrm{reset}} = 0.15$ V in Fig. 6.9, and (2) the correlation between R and I_{C} in Fig. 6.5a depends on the latest steps of the filament evolution, largely in the growth regime when the model in Eq. (6.10) already applies. Therefore, to understand and account for the set/reset characteristics in Figs. 6.5 and 6.7, the growth model provides sufficient physical accuracy.

The reset process can be modeled by Eq. (6.10) with a simple change in the sign of the preexponential coefficient A, to describe a decrease in size of the CF, instead of a growth. Such a reset model fully accounts for the Joule heating and filament evolution in the initial stages of the reset process, when there is a continuous CF connecting the two electrodes. On the other hand, the model may fail in describing the later stages of reset, when the gap opening results in a change in the thermal and electrical description of the filament. However, the model fully accounts for the calculation of parameters V_{reset} and I_{reset} in Fig. 6.5b, which are responsible for the initiation of the reset process.

6.7 Simulation Results

Figure 6.13 shows measured and calculated resistance and voltage across the cell for different load resistance $R_{\mathrm{L}} = 1$ kΩ (a, d), 2.2 kΩ (b, e), 5 kΩ (c, f). Data are the same as in Fig. 6.7, while calculations were performed with Eq. (6.10) assuming $E_{\mathrm{A}} = 1.2$ eV, $A = 1$ ms^{-1} and $\alpha = 0.3$ [74]. The value chosen for the activation energy is in agreement with the energy barrier for diffusion and migration observed in various oxide systems, including NiO [62, 63, 76] and Gd-doped MoO [80]. Such a value should be viewed as the energy barrier for ion migration according to the schematic of Fig. 6.12. The value used for the barrier lowering coefficient α is

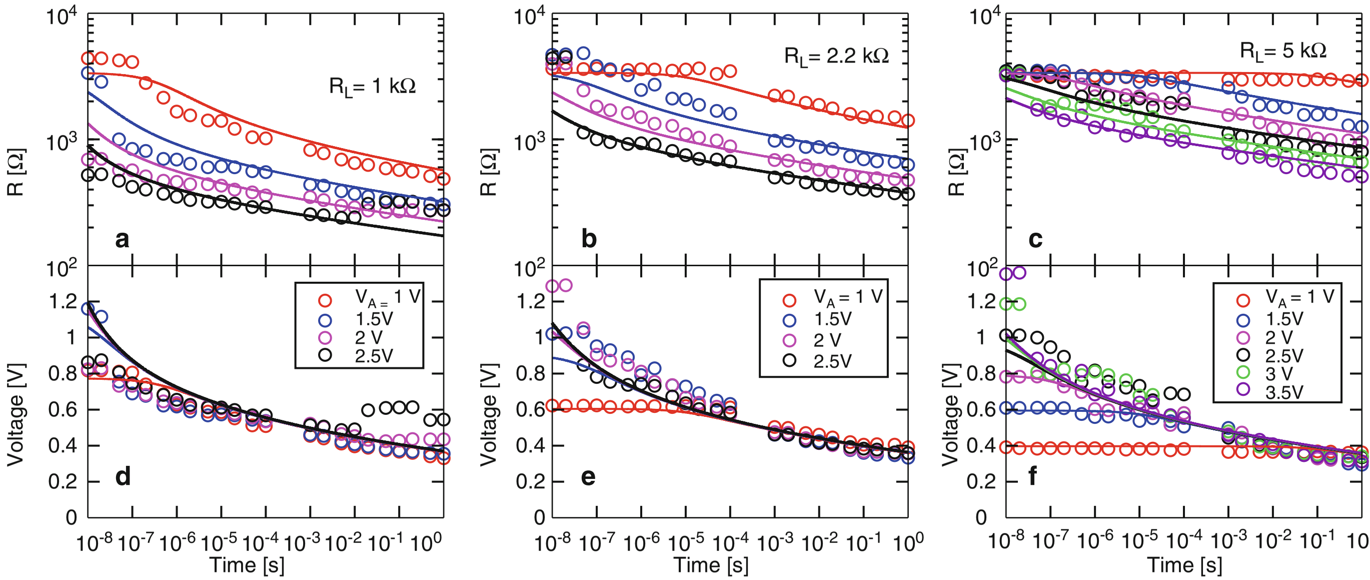


Fig. 6.13 Measured and calculated R and V for increasing load resistance $R_L = 1\ \mathrm{k}\Omega$ (**a**, **d**), 2.2 kΩ (**b**, **e**), and 5 kΩ (**c**, **f**). Calculations were performed using Eq. (6.10) with parameters $E_{A0} = 1.4$ eV, $\alpha = 0.3$, and $A = 1\ \mathrm{m\ s}^{-1}$. The model accounts for (1) the voltage acceleration of filament growth, (2) the dependence on R_L due to current limitation, and (3) the universal time evolution of V, due to the exponential dependence of filament growth on V. Reprinted with permission from [74] (© 2011 IEEE)

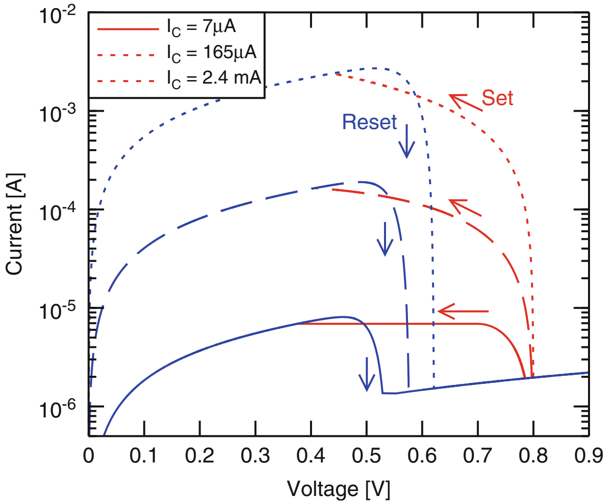


Fig. 6.14 Calculated *I*–*V* characteristics for set and reset at increasing I_C = 7 μA, 165 μA, and 2.4 mA. A 1T1R structure was assumed in the calculations, with a constant applied voltage V_A = 0.8 V during set and a voltage sweep during reset. Reprinted with permission from [61] (© 2011 IEEE)

similar to previous analysis [75, 79]. The choice of each of these values will be later sustained by individual comparisons with data. The calculations in the figure account very well for experimental data, in terms of voltage acceleration of the set process, where the increase of V_A results in an increase of the growth speed revealed by the decrease of the resistance, and of the compliance effect, where the increase of R_L results in a stronger limitation of the current, thus limiting the final size of the CF. In particular, note that the model accounts for the universal dependence of *V* on *t* in Fig. 6.13d–f, irrespective of the applied voltage and load resistance. This is because the growth rate in Eq. (6.10) is a strong function of *V*, thus enabling a regulation of the voltage across to the device thanks to the negative feedback loop, where any filament growth results in a decrease of voltage.

Figure 6.14 shows the calculated *I*–*V* curve during set and reset for increasing values of I_C = 7 μA, 166 μA and 2.4 mA. Simulations were performed assuming a 1T1R structure, where the saturated current in the MOS transistor was given by $I_{sat} = KV_{OD}^2/2$, with $K = 700$ μA cm^{-2}. V_{OD} represents the overdrive voltage, namely, the gate voltage minus the threshold voltage of the MOS transistor. The current compliance is thus given by the saturated MOS current and was changed in the simulations by changing the gate voltage at the MOS transistor. A voltage V_A = 0.8 V was applied during set from a reset state consisting of a high resistance state with an initial filament of ϕ_N = 0.5 nm diameter. During set, the current increases due to filament growth and the consequent decrease of resistance. Once the saturated current is approached, the voltage across the device decreases similar to Eq. (6.3) (although no load resistance can properly be defined in the case of

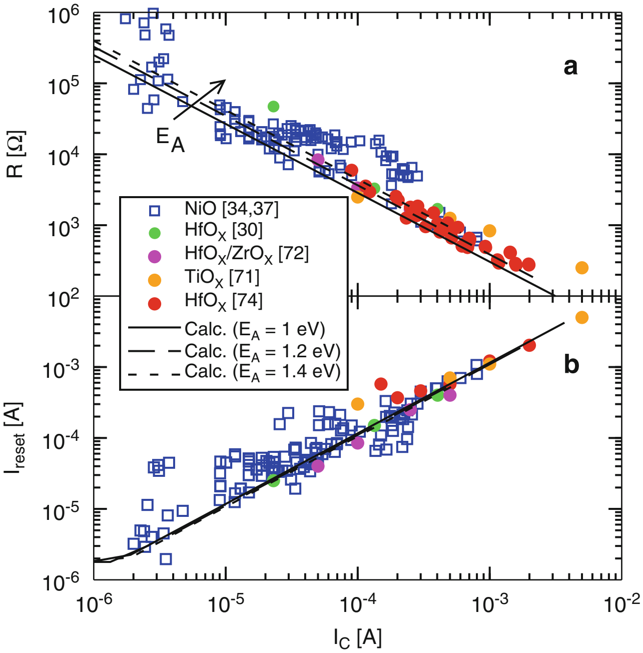


Fig. 6.15 Measured and calculated R in the set state (**a**) and reset current I_{reset} (**b**) as a function of I_C, for both unipolar and bipolar switching and for several metal oxides, including NiO [21, 34, 36, 37], TiO_x [71], HfO_x [30, 74], and HfO_x–ZrO_x [72]. Calculations by the filament growth model accounts for Eqs. (6.1) and (6.2) describing the I_C dependence of R and I_{reset}, respectively. The constant voltage V_C is due to the voltage regulation effect during set under a limited current compliance, as shown in Figs. 6.13 and 6.14. The similar values for I_{reset} and I_C are due to the common nature of set and reset mechanisms, driven by the same activation energy E_{A0}. The universal R and I_{reset} behaviors are due to the weak dependence on E_{A0}, as shown by calculations for variable energy barrier. Reprinted with permission from [61] (© 2011 IEEE)

nonlinear MOS characteristics). After 1 s, which was the maximum set time in the calculation, the voltage reached a value V_C of about 0.4 V, and thus the final resistance obeys Eq. (6.1). During reset the MOS transistor was biased to a large conductivity to allow negligible series resistance in the 1T1R. The device current increases according to the I–V curves in the final set state until reset takes place at V_{reset}, which is slightly larger than V_C. As a result, the reset current I_{reset} is only slightly larger than I_C, thus satisfying Eq. (6.2). Note the abrupt current drop in the calculations, which is due to the rapid decrease of CF size and the consequent increase of resistance as the temperature reaches the critical value for the onset of ion migration.

Figure 6.15 shows the measured and calculated R (a) and I_{reset} (b) as a function of I_C, for several oxide-based RRAM in the literature [21, 30, 34, 36, 37, 71, 72]. Calculations were done assuming a total time of 1 s during set and a sweep rate of 2 V s^{-1} during reset, corresponding to DC switching. Calculations in Fig. 6.15a, b satisfy Eqs. (6.1) and (6.2), respectively, for set and reset, as a result of the voltage-controlled set/reset kinetics. Calculations were done at variable $E_A = 0.9$, 1.2 and 1.5 eV, while A and α were kept unchanged. A decrease of E_A leads to a weak decrease in V_C: This is because the energy barrier decreases enhances ion migration, thus a lower temperature (hence voltage) is needed to sustain filament growth at any given time. The voltage thus readjusts to a smaller value during set under the negative feedback condition of Eq. (6.3). However, the change of V_C is relatively small, about 0.03 V for a 0.1 eV change of E_A in the range considered. Such a small sensitivity of set characteristic to E_A might explain why most oxide materials display the same V_C in Fig. 6.15a. Note also that the similarity between I_{reset} and I_C, namely $\eta \approx 1$ in Eq. (6.2), can be attributed to the fact that set and reset processes rely on the same physical mechanism, namely temperature and field activated ion migration. This ensures that the same voltage (hence temperature) is needed to activate migration during either set or reset, irrespective of the polarity, thus causing $V_{reset} = V_C$. As a consequence, the reset I_{reset} is also approximately equal to I_C, as already stated in Eq. (6.2) where $\eta \approx 1$.

To better understand the set mechanism, and in particular the universal time evolution of the voltage across the cell during set, one may consider rewriting Eq. (6.10) in the following way:

$$\mathrm{d}t = A^{-1}\mathrm{d}\phi \mathrm{e}^{\frac{E_A}{k\left(T_0 + R_{th}/RV^2\right)}} = A^{-1}\mathrm{d}\phi \mathrm{e}^{\frac{E_A}{kT_0\left(1+\beta I_C^2/\phi^4\right)}}, \tag{6.11}$$

where E_A has been assumed a constant for simplicity, an ideal current limitation to a constant $I_C = V/R$ was considered and the parameter β has been introduced, given by:

$$\beta = \frac{2\rho L^2}{\pi^2 k_{th} T_0}. \tag{6.12}$$

The integration of Eq. (6.11) leads to:

$$t_P = \int_{\phi_N}^{\phi_{set}} \mathrm{d}\phi A^{-1} \mathrm{e}^{\frac{E_A}{kT_0\left(1+\beta I_C^2/\phi^4\right)}} = \int_{\phi_N}^{\phi_{set}} f(\phi)\mathrm{d}\phi, \tag{6.13}$$

where t_P is the total duration of the set pulse, $f(\phi)$ is the function within the integral, and ϕ_{set} is the final diameter of the filament at the end of the set transition. The function $f(\phi)$ in Eq. (6.13) cannot be integrated analytically. Figure 6.16 shows the calculated function f as a function of ϕ (a) and the corresponding integral, namely the duration of the set time (b), for three values of the current compliance $I_C = 1$, 10 and 100 μA. Both ϕ and its integral are steeply increasing functions of the diameter, which indicates that the filament growth under constant I_C extends over several decades of time with relatively little increase of diameter. The integral of Eq. (6.13) can be found by interpolation of the calculated curves at the total set time t_P, and the

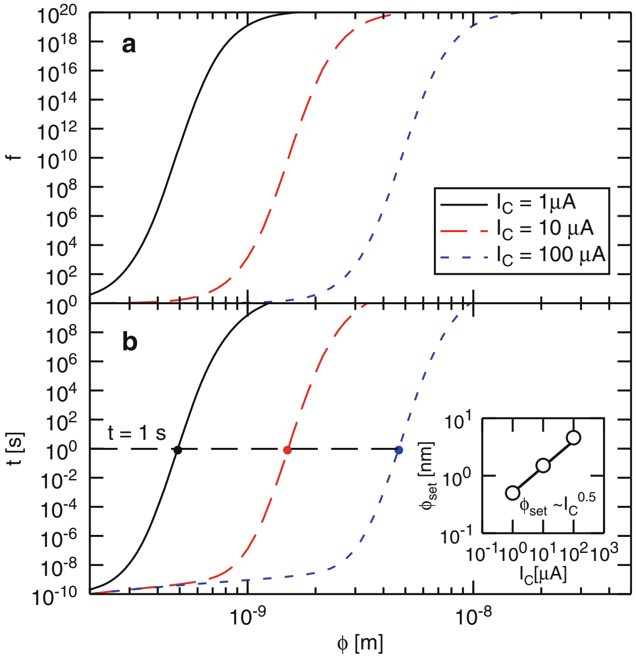


Fig. 6.16 Calculated function f in the integral of Eq. (6.13) (**a**) and time t resulting from the integration of the same equation (**b**) for increasing $I_C = 1$, 10, and 100 μA. The *curves* of calculated t for a set time $t_P = 1$ s indicate the final filament size after set transition, as shown in the *inset* as a function of I_C. The proportionality $\phi_{set} \propto {I_C}^{0.5}$ is consistent with $R \propto {I_C}^{-1}$, thus providing a physical basis for the empirical formula in Eq. (6.1)

result is shown for $t_P = 1$ s, corresponding to a typical DC switching experiment. In all cases, the resulting ϕ_{set} is found in the steep region of the integral of Fig. 6.16b. From inspection of Eq. (6.13), one may expect that the condition $t_P = 1$ s is satisfied by a critical value of the term $C^2 = \beta {I_C}^2/\phi^4$ in the denominator of the exponent. This is demonstrated by the inset of Fig. 6.16, showing the extracted ϕ_{set} as a function of I_C and showing that ϕ_{set} is proportional to ${I_C}^{0.5}$. Note that latter can be viewed as $R \propto \phi^{-2} \propto {I_C}^{-1}$, namely Eq. (6.1). The physical meaning of the parameter $C = \beta^{0.5} I_C/\phi^2$ controlling the exponent in the integral of Eq. (6.13) can be understood by a simple elaboration, showing that:

$$C = \frac{\beta^{0.5} I_C}{\phi^2} = \frac{R I_C}{\sqrt{2k_{th}\rho T_0}} = \frac{V}{\sqrt{2k_{th}\rho T_0}}, \tag{6.14}$$

namely C is proportional to V through a constant $(2k_{th}\rho T_0)^{-1}$. Therefore, Eq. (6.13) indicates that there is a correspondence between V across the cell and time during the set transition. This is shown in Fig. 6.17, showing $C \propto V$ as a function of time for the three values of I_C in Fig. 6.16: C (hence V) follows a universal function of time, irrespective of the current compliance, in agreement with results shown in Fig. 6.13.

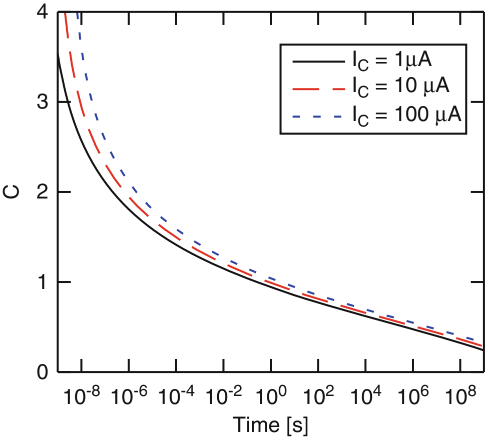


Fig. 6.17 Calculated parameter $C = \beta^{0.5} I_C \phi^{-2}$ as a function of time according to Eq. (6.14), for increasing I_C. The independence from I_C accounts for the universal time evolution of voltage in Fig. 6.13

6.8 Switching Time and Energy

Switching speed and switching energy are among the most important properties defining the performance of a new device for low power, high speed integrated circuits. Figure 6.18 shows the calculated reset time t_{reset} as a function of V (a) and as a function of $1/kT$ (b). The figure also shows experimental data for set and reset times (same data as in Fig. 6.9). The same parameter values for E_A, A, and α in Fig. 6.13 were used, while three different CF sizes were used, corresponding to $R = 220\ \Omega$, 10 kΩ, and 400 kΩ. The model accounts very well for the switching times in both the voltage and Arrhenius plots. In particular, the Arrhenius behavior, which is a key feature in the filament growth model, is clearly indicated by data aligning on a straight line in Fig. 6.13b. A more detailed analysis reveals that the data and calculations in Fig. 6.13b are slightly nonlinear: this can be understood by the voltage-induced barrier lowering, which results in a decrease of E_A (i.e., a flattening of the curve in the Arrhenius plot) for increasing voltage, hence decreasing $1/kT$. These results also provide a direct way to extract the microscopic parameters for ion migration from experimental data, namely the activation energy E_A can be extracted from the slope of data in the Arrhenius plot and the barrier lowering coefficient α can be extracted from the curvature of data in the Arrhenius pot. The figure also compares calculations for different initial sizes of the filament, showing that a smaller CF requires a shorter time for reset. Given the symmetry between set and reset, the same conclusion might be drawn for the set transition, provided that set is conducted at constant voltage as assumed for reset in Fig. 6.13. No dependence on final filament size is instead expected when variable CF sizes are achieved through different compliance levels, such as in Figs. 6.4 and 6.7. A reduction of set/reset times by roughly a factor 40 is expected for an increase of resistance from 220 Ω to 400 kΩ. This factor reflects the difference in the initial diameter ϕ_{set} of the filament, which can be estimated as $(400\ \mathrm{k}\Omega/220\ \Omega)^{0.5} \approx 40$.

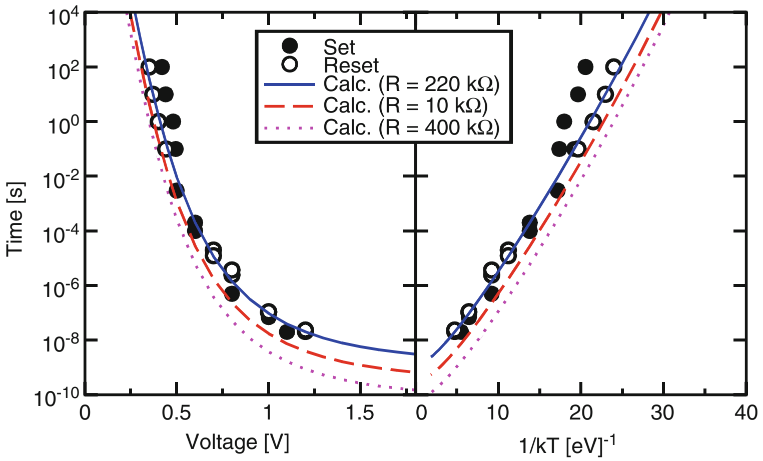


Fig. 6.18 Measured set and reset times, as a function of V (**a**) and $1/kT$ (**b**), where T is the local temperature at the switching location. Calculations of the reset time are also shown for increasing $R = 220\ \Omega$, 10 kΩ, and 400 kΩ. Note that the curve in the Arrhenius plot is not perfectly linear, as a result of the voltage-induced energy barrier lowering. Adapted from [74] (© 2011 IEEE)

In fact, according to the filament growth model, the reset time can be given by the formula [76]:

$$t_{\mathrm{reset}} = \frac{\phi_{\mathrm{set}}}{2v_{\mathrm{G}}}, \tag{6.15}$$

where v_G is an effective growth velocity. Eq. (6.15) indicates that, for any given voltage, the reset time should be proportional to the initial filament size ϕ_{set}.

Figure 6.19 shows the calculated reset energy E_{reset} as a function of voltage for the three resistance values used in Fig. 6.18. Data for an initial resistance of 220 Ω are also shown for reference. The energy was calculated by the integral:

$$E_{\mathrm{reset}} = \int_{\mathrm{reset}} VI\mathrm{d}t = \int_{\mathrm{reset}} \frac{V^2}{R}\mathrm{d}t, \tag{6.16}$$

that is the integral of dissipated power in time during the reset transition. Note that only the energy dissipation in the device is considered in Eq. (6.16), while more energy could be dissipated at the series transistor/resistance or at the select device. The set energy E_{set} dissipated in the device is equal to E_{reset} for a constant voltage set, since set is just a reversed reset process in this case. For an I_{C}-controlled set, such as in Fig. 6.14, the E_{set} is expected to be even smaller than E_{reset}, because, although V could be higher to provide the nucleation overvoltage, the transition becomes faster at high voltage according to the exponential relationship in Eq. (6.10), therefore reducing the product $V^2\mathrm{d}t$ in Eq. (6.16).

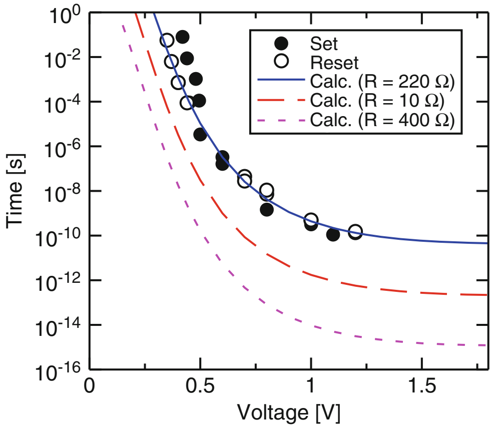


Fig. 6.19 Measured and calculated set/reset energy as a function of V for $R = 220\ \Omega$, 10 kΩ, and 400 kΩ. The energy largely decreases for increasing V, due to the switching times decreasing almost exponentially in Fig. 6.18, and for increasing R, as a result of the decreasing I in Eq. (6.16). Adapted from [74] (© 2011 IEEE)

Results in Fig. 6.19 suggest two ways to efficiently reduce the reset energy in RRAM: First, E_{reset} decreases significantly for increasing voltage, as a result of the almost exponential enhancement of transition speed with voltage in Eq. (6.10). Second and most importantly, reducing the size of the CF results in a remarkable reduction of E_{reset} at a given voltage, thanks to the increase of R in Eq. (6.16). From the figure, reset energies of the order of 10 fJ can be achieved by reset at about 1 V on a RRAM device with 400 kΩ resistance, corresponding to 2.5 μA reset current and 5 ns switching.

6.9 Scaling Challenges

From the results in Figs. 6.18 and 6.19, CF size control and reduction appears as an effective way to reduce both transition time and energy consumption in RRAM. To achieve such large CF resistances, however, one should ensure a sufficiently high resistance in the reset state, too. In this perspective, having a large resistance window available in the RRAM device is a key requirement, since it allows writing extremely small CFs by taking advantage of the improved programming speed and reduced energy consumption.

The reduction of the CF size has another important impact on the scaling of memory arrays with crossbar architecture. In fact, one of the most attractive features of RRAM is the ability to organize the memory array in a crossbar circuit, where each memory cell occupies an area of only $4F^2$ [4, 50, 54]. However, such architecture is prone to interference during program and read. For instance, application of a voltage across a cell during read results in a sneak-through current through cell belonging to the same row or column: this may result in a major misinterpretation of the bit status, when the cell to be read is in the high resistance state [44, 49, 50]. To solve this issue, each memory cell must be accompanied by a nonlinear selector, e.g., a rectifier diode [49–51]. Several types of select devices

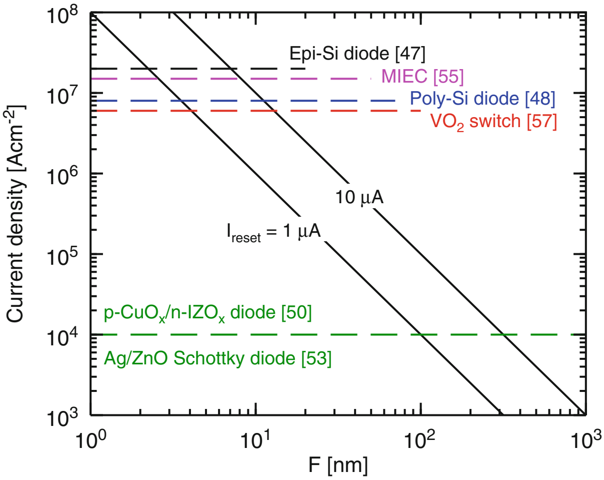


Fig. 6.20 Reset current density as a function of the technology node F, assuming a constant reset current $I_{\text{reset}} = 1$ and 10 μA. The reset current increases for decreasing F according to $I_{\text{reset}} \propto F^{-2}$. The current density reported in literature for Si and non-Si diodes is also shown for comparison [47–53, 55, 56]. Reprinted with permission from [37] (© 2011 Elsevier, Ltd.)

have been proposed so far, including monocrystalline silicon diodes [47], polycrystalline silicon diodes [48], oxide unipolar heterostructure diodes [49–52], Schottky diodes [53], mixed ionic–electronic conduction diodes [55, 56], and VO_2 threshold switches [51, 57]. For the purpose of diode screening for selector applications, three critical requirements should be considered: First, the diode should be available in the back-end of the line, to allow the stacking of several memory layers. Si-based selectors may hardly be compatible with such requirement due to the high temperatures needed for deposition and doping diffusion/activation. The second key requirement is a sufficient ON/OFF current ratio of the selector, to allow for sufficient blocking of the current through unselected devices during read. Finally, an important requirement is the capability to supply sufficient current during program, to allow for the set/reset of the memory element. Such a requirement seems the most hard to be met: Fig. 6.20 shows the current density at about 2 V of reported select elements, compared to the reset current density as a function of the device size F [37]. Note that the required current density increases for decreasing F, and thus with the down scaling of the memory cell. Therefore, to meet the reset current requirement in future technology nodes, one should ensure (1) satisfactory current density in the select element and (2) sufficiently small reset current $I_{\text{reset}} = V_{\text{reset}}/R$, and thus sufficiently small CFs. From this standpoint, the size reduction of filaments in oxide RRAM is essential.

In view of CF scaling, it should also be mentioned that CF size reduction might conflict with reliability requirements, namely data retention and noise issues. Data retention at elevated temperature was in fact shown to strongly depend on the CF

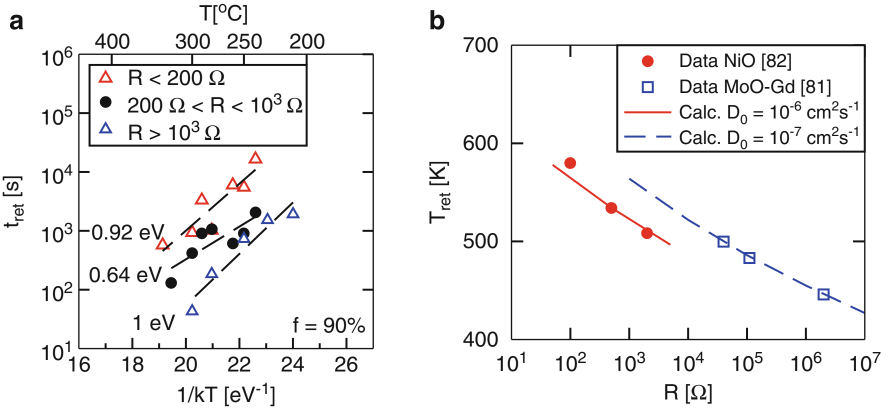


Fig. 6.21 Measured retention times at 10 % percentile [81] (**a**) and measured/calculated retention temperatures as a function of initial resistance [80, 81] (**b**). Three different initial resistances were used in (**a**), namely less than 200 Ω, between 200 and 1,000 Ω, and above 1,000 Ω. The retention times obey the Arrhenius law, however $t_{\rm ret}$ decreases for increasing R, as a result of size-dependent oxidation [81]. The retention temperature change at 1,000 s can be reproduced by the analytical reset model of Eq. (6.18). Reprinted with permission from [81] and [82] (© 2011 IEEE and ECS)

resistance, since smaller filaments with higher resistances were found to display shorter retention time according to the formula [80, 81]:

$$t_{\rm ret} = \frac{\phi_{\rm set}^2}{D_0} e^{\frac{E_{\rm A}}{kT}} \tag{6.17}$$

where D_0 is a diffusivity factor. Equation (6.17) is based on a diffusion model, where the diffusion of conductive atoms from a CF is driven by the concentration gradient, and thus critically depends on CF size [77]. Figure 6.21a shows the Arrhenius plot of retention times for a failure rate of 10 %, i.e., the retention time of one device out of 10 devices on the average was shorter than $t_{\rm ret}$ in the figure. Data are reported for a cell initially programmed within three different resistance ranges, namely below 200 Ω, between 200 Ω and 1 kΩ, and above 1 kΩ. Data retention time decreases for increasing resistance, and thus decreasing size of the CF. Figure 6.21b shows the retention temperature corresponding to $t_{\rm ret} = 10^3$ s as a function of initial resistance, from Fig. 6.21a [82] and from MoO-Gd RRAM devices [80]. The retention temperature $T_{\rm ret}$ decreases with R according to the formula:

$$T_{\rm ret} = \frac{E_{\rm A}}{k \log \frac{D_0 t_{\rm ret}}{\phi_{\rm set}^2}} \approx \frac{E_{\rm A}}{k \log \frac{\pi R D_0 t_{\rm ret}}{4\rho L}}, \tag{6.18}$$

where the ohmic approximation for R was used. Calculations based on Eq. (6.17) are also shown in the figure, supporting the size-dependent retention model.

Another potential concern for small CFs is the fluctuation of the current during read, which might be induced by surface charging/discharging effects due to surface rearrangement [83]. The relative amplitude of random telegraph noise (RTN) has been shown to increase for decreasing CF size, as a result of the stronger impact of surface conduction [83]. From this viewpoint, the trade-off between reset time/energy reduction and reliability must be carefully assessed.

6.10 Conclusions and Outlook

RRAM is the strongest candidate for high-density nonvolatile memory technologies below the 10 nm node. The RRAM technology offers several key advantages, such as switching speed, low voltage operation, good retention and endurance reliability, and low cost. However, the scalability of RRAM is still under debate, due to the lack of understanding and physically based models for the switching mechanisms. This review provides an overview of the recent progress on the physical interpretation and modeling of the oxide-based bipolar switching RRAM. The switching mechanism has been discussed based on experimental results regarding the time evolution of resistance and voltage across the cell, revealing the key role of voltage as the controlling parameter for the switching characteristic. The voltage dependence of switching times has provided evidence for an Arrhenius law, revealing the temperature-activated nature of the switching process and the key role of voltage-driven Joule heating at the localized filament. It has been thus concluded that set/reset processes of bipolar switching RRAM rely on thermally activated ion migration driven by the electric field. Based on this physical interpretation, a model has been developed to describe the filament growth and dissolution during set and reset, respectively. The filament grows/dissolves through ion migration in the direction of the field, and such process is strongly accelerated by the local temperature, which largely increases through Joule heating. The model was tested against the time dependence of set dynamics, the set/reset parameters as a function of current compliance, and the set/reset times. Finally, the consequences of this new understanding and modeling in terms of scaling have been discussed. In particular, RRAM technology will face a severe challenge in matching, on the one hand the requirement of reducing time, energy, and space within the chip (i.e., cost), and, on the other hand, the current density limits of select devices and size-dependent reliability issues of the RRAM device.

Acknowledgments The author would like to acknowledge many students and colleagues, in particular F. Nardi, S. Balatti, C. Cagli, and A. L. Lacaita. Many fruitful discussions with K. Min, D. Kau, G. Spadini, R. Bruchhaus, R. Waser, S. Spiga, D. Deleruyelle, C. Muller, and M. Kozicki are also gratefully acknowledged. This work was supported in part by Fondazione Cariplo (Grant number 2010-0500) and Intel (Project 55887).

References

1. R. Waser, M. Aono, Nanoionics-based resistive switching memories. Nat. Mater. **6**, 833–840 (2007)
2. A. Sawa, Resistive switching in transition metal oxides. Mater. Today **11**, 28–36 (2008)
3. R. Waser, R. Dittmann, G. Staikov, K. Szot, Redox-based resistive switching memories—nanoionic mechanisms, prospects, and challenges. Adv. Mater. **21**, 2632–2663 (2009)
4. D. Ielmini, R. Bruchhaus, R. Waser, Thermochemical resistive switching: Materials, mechanisms and scaling projections. Phase Transit. **84**, 570–602 (2011)
5. H. Akinaga, H. Shima, Resistive random access memory (ReRAM) based on metal oxides. Proc. IEEE **98**, 2237–2251 (2010)
6. M.N. Kozicki, M. Park, M. Mitkova, Nanoscale memory elements based on solid-state electrolytes. IEEE Trans. Nanotechnol. **4**, 331–338 (2005)
7. M. Kund, G. Beitel, C.-U. Pinnow, T. Rohr, J. Schumann, R. Symanczyk, K. Ufert, G. Muller, Conductive bridging RAM (CBRAM): An emerging non-volatile memory technology scalable to sub 20 nm. IEDM Tech. Dig. 754–757 (2005)
8. S. Lombardo, J.H. Stathis, B.P. Linder, K.L. Pey, F. Palumbo, C.H. Tung, Dielectric breakdown mechanisms in gate oxides. J. Appl. Phys. **98**, 121301 (2005)
9. X. Li, C.H. Tung, K.L. Pey, The nature of dielectric breakdown. Appl. Phys. Lett. **93**, 072903 (2008)
10. X. Li, C.H. Tung, K.L. Pey, V.L. Lo, The chemistry of gate dielectric breakdown. IEDM Tech. Dig. 779–782 (2008)
11. G.-S. Park, X.-S. Li, D.-C. Kim, R.-J. Jung, M.-J. Lee, S. Seo, Observation of electric-field induced Ni filament channels in polycrystalline NiO_x film. Appl. Phys. Lett. **91**, 222103 (2007)
12. D.-H. Kwon, K.M. Kim, J.H. Jang, J.M. Jeon, M.H. Lee, G.H. Kim, X.-S. Li, G.-S. Park, B. Lee, S. Han, M. Kim, C.S. Hwang, Atomic structure of conducting nanofilaments in TiO_2 resistive switching memory. Nat. Nanotechnol. **5**, 148–153 (2010)
13. J.P. Strachan, M.D. Pickett, J.J. Yang, S. Aloni, A.L.D. Kilcoyne, G. Medeiros-Ribeiro, R.S. Williams, Direct identification of the conducting channels in a functioning memristive device. Adv. Mater. **22**, 3573 (2010)
14. C. Schindler, S.C.P. Thermadam, R. Waser, M.N. Kozicki, Bipolar and unipolar resistive switching in Cu-doped SiO_2. IEEE Trans. Electron Devices **54**, 2762–2768 (2007)
15. C. Schindler, Resistive switching in electrochemical metallization memory cells. PhD Thesis, Rheinisch-Westfalischen Technischen Hochschule Aachen, 2009
16. J.F. Gibbons, W.E. Beadle, Switching properties of thin NiO films. Solid State Electron. **7**, 785–790 (1964)
17. F. Argall, Switching phenomena in titanium oxide thin films. Solid State Electron. **11**, 535–541 (1968)
18. W.R. Hiatt, T.W. Hickmott, Bistable switching in niobium oxide diodes. Appl. Phys. Lett. **6**, 106–108 (1965)
19. A. Beck, J.G. Bednorz, C. Gerber, C. Rossel, D. Widmer, Reproducible switching effect in thin oxide films for memory applications. Appl. Phys. Lett. **77**, 139–141 (2000)
20. Y. Watanabe, J.G. Bednorz, A. Bietsch, C. Gerber, D. Widmer, A. Beck, S.J. Wind, Current-driven insulator–conductor transition and nonvolatile memory in chromium-doped $SrTiO_3$ single crystals. Appl. Phys. Lett. **78**, 3738–3740 (2001)
21. S. Seo, M.J. Lee, D.H. Seo, E.J. Jeoung, D.-S. Suh, Y.S. Joung, I.K. Yoo, I.R. Hwang, S.H. Kim, I.S. Byun, J.-S. Kim, J.S. Choi, B.H. Park, Reproducible resistance switching in polycrystalline NiO films. Appl. Phys. Lett. **85**, 5655–5657 (2004)
22. I.G. Baek, M.S. Lee, S. Seo, M.J. Lee, D.H. Seo, D.-S. Suh, J.C. Park, S.O. Park, H.S. Kim, I.K. Yoo, U.-I. Chung, J.T. Moon, Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses. IEDM Tech. Dig. 587–590 (2004)
23. S.R. Lee, K. Char, D.C. Kim, R. Jung, S. Seo, X.S. Li, G.-S. Park, I.K. Yoo, Resistive memory switching in epitaxially grown NiO. Appl. Phys. Lett. **91**, 202115 (2007)

24. H. Shima, F. Takano, H. Akinaga, Resistance switching in the metal deficient-type oxides: NiO and CoO. Appl. Phys. Lett. **91**, 012901 (2007)
25. B.J. Choi, D.S. Jeong, S.K. Kim, C. Rohde, S. Choi, J.H. Oh, H.J. Kim, C.S. Hwang, K. Szot, R. Waser, B. Reichenberg, S. Tiedke, Resistive switching mechanism of TiO_2 thin films grown by atomic-layer deposition. J. Appl. Phys. **98**, 033715 (2005)
26. D.B. Strukov, G.S. Snider, D.R. Stewart, R.S. Williams, Nature **443**, 80–83 (2008)
27. W. Wang, S. Fujita, S.S. Wong, RESET mechanism of TiO_x resistance-change memory device. IEEE Electron Device Lett. **30**, 733–735 (2009)
28. K.M. Kim, C.S. Hwang, The conical shape filament growth model in unipolar resistance switching of TiO_2 thin film. Appl. Phys. Lett. **94**, 122109 (2009)
29. T.-N. Fang, S. Kaza, S. Haddad, A. Chen, Y.-C. Wu, Z. Lan, S. Avanzino, D. Liao, C. Gopalan, S. Choi, S. Mahdavi, M. Buynoski, Y. Lin, C. Marrian, C. Bill, M. VanBuskirk, M. Taguchi, Erase mechanism for copper oxide resistive switching memory cells with nickel electrode. IEDM Tech. Dig. 789–792 (2006)
30. H.Y. Lee, P.S. Chen, T.Y. Wu, Y.S. Chen, C.C. Wang, P.J. Tzeng, C.H. Lin, F. Chen, C.H. Lien, M.-J. Tsai, Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO_2 based RRAM. IEDM Tech. Dig. 297–300 (2008)
31. Y.S. Chen, H.Y. Lee, P.S. Chen, P.Y. Gu, C.W. Chen, W.P. Lin, W.H. Liu, Y.Y. Hsu, S.S. Sheu, P.C. Chiang, W.S. Chen, F.T. Chen, C.H. Lien, M.-J. Tsai, Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity. IEDM Tech. Dig. 105–108 (2009)
32. J.J. Yang, M.-X. Zhang, J.P. Strachan, F. Miao, M.D. Pickett, R.D. Kelley, G. Medeiros-Ribeiro, R.S. Williams, High switching endurance in TaO_x memristive devices. Appl. Phys. Lett. **97**, 232102 (2010)
33. M.-J. Lee, C.B. Lee, D. Lee, S.R. Lee, M. Chang, J.H. Hur, Y.-B. Kim, C.-J. Kim, D.H. Seo, S. Seo, U.-I. Chung, I.-K. Yoo, K. Kim, A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structures. Nat. Mater. **10**, 625–630 (2011)
34. K. Tsunoda, K. Kinoshita, H. Noshiro, Y. Yamazaki, T. Iizuka, Y. Ito, A. Takahashi, A. Okano, Y. Sato, T. Fukano, M. Aoki, Y. Sugiyama, IEDM Tech. Dig. 767–770 (2007)
35. Y. Sato, K. Tsunoda, K. Kinoshita, H. Noshiro, M. Aoki, Y. Sugiyama, Sub-100 μA reset current of nickel oxide resistive memory through control of filamentary conductance by current limit of MOSFET. IEEE Trans. Electron Devices **55**, 1185–1191 (2008)
36. K. Kinoshita, K. Tsunoda, Y. Sato, H. Noshiro, S. Yagaki, M. Aoki, Y. Sugiyama, Reduction in the reset current in a resistive random access memory consisting of NiO_x brought about by reducing a parasitic capacitance. Appl. Phys. Lett. **93**, 033506 (2008)
37. F. Nardi, D. Ielmini, C. Cagli, S. Spiga, M. Fanciulli, L. Goux, D.J. Wouters, Control of filament size and reduction of reset current below 10 μA in NiO resistance switching memories. Solid State Electron. **58**, 42–47 (2011)
38. Y. Wu, B. Lee, H.-S.P. Wong, Al_2O_3-based RRAM using atomic layer deposition (ALD) with 1 μA reset current. IEEE Electron Device Lett. **31**, 1449–1441 (2010)
39. C.H. Cheng, A. Chin, F.S. Yeh, High performance ultra-low energy RRAM with good retention and endurance. IEDM Tech. Dig. 448–491 (2010)
40. C.H. Ho, C.-L. Hsu, C.-C. Chen, J.-T. Liu, C.-S. Wu, C.-C. Huang, C. Hu, F.-L. Yang, 9 nm Half-pitch functional resistive memory cell with < 1 μA programming current using thermally oxidized sub-stoichiometric WO_x film. IEDM Tech. Dig. 436–439 (2010)
41. D. Ielmini, C. Cagli, F. Nardi, Resistance transition in metal oxides induced by electronic threshold switching. Appl. Phys. Lett. **94**, 063511 (2009)
42. H.Y. Lee, Y.S. Chen, P.S. Chen, P.Y. Gu, Y.Y. Hsu, S.M. Wang, W.H. Liu, C.H. Tsai, S.S. Sheu, P.C. Chiang, W.P. Lin, C.H. Lin, W.S. Chen, F.T. Chen, C.H. Lien, M.-J. Tsai, Evidence and solution of over-RESET problem for HfO_x based resistive memory with sub-ns switching speed and high endurance. IEDM Tech. Dig. 460–463 (2010)

43. F. Nardi, S. Balatti, S. Larentis, D. Ielmini, Complementary switching in metal oxides: Toward diode-less crossbar RRAMs. IEDM Tech. Dig. 709–712 (2011)
44. E. Linn, R. Rosezin, C. Kügeler, R. Waser, Complementary resistive switches for passive nanocrossbar memories. Nat. Mater. **9**, 403–406 (2010)
45. D. Ielmini, F. Nardi, C. Cagli, Universal reset characteristics of unipolar and bipolar metal-oxide RRAM. IEEE Trans. Electron Devices **58**, 3246–3253 (2011)
46. R. Rosezin, E. Linn, L. Nielen, C. Kügeler, R. Bruchhaus, R. Waser, Integrated complementary resistive switches for passive high-density nanocrossbar arrays. IEEE Electron Device Lett. **32**, 191–193 (2011)
47. J.H. Oh, J.H. Park, Y.S. Lim, H.S. Lim, Y.T. Oh, J.S. Kim, J.M. Shin, J.H. Park, Y.J. Song, K.C. Ryoo, D.W. Lim, S.S. Park, J.I. Kim, J.H. Kim, J. Yu, F. Yeung, C.W. Jeong, J.H. Kong, D.H. Kang, G.H. Koh, G.T. Jeong, H.S. Jeong, K. Kim, Full integration of highly manufacturable 512Mb PRAM based on 90 nm technology. IEDM Tech. Dig. 515–518 (2006)
48. Y. Sasago, M. Kinoshita, T. Morikawa, K. Kurotsuchi, S. Hanzawa, T. Mine, A. Shima, Y. Fujisaki, H. Kume, H. Moriya, N. Takaura, K. Torii, Cross-point phase change memory with $4F^2$ cell size driven by low-contact resistivity poly-Si diode. Symp. VLSI Tech. Dig. 24–25 (2009)
49. I.G. Baek, D.C. Kim, M.J. Lee, H.-J. Kim, E.K. Yim, M.S. Lee, J.E. Lee, S.E. Ahn, S. Seo, J.H. Lee, J.C. Park, Y.K. Cha, S.O. Park, H.S. Kim, I.K. Yoo, U.-I. Chung, J.T. Moon, B.I. Ryu, Multi-layer cross-point binary oxide resistive memory (OxRRAM) for post-NAND storage application. IEDM Tech. Dig. 750–753 (2005)
50. M.-J. Lee, Y. Park, B.-S. Kang, S.-E. Ahn, C. Lee, K. Kim, W. Xianyu, G. Stefanovich, J.-H. Lee, S.-J. Chung, Y.-H. Kim, C.-S. Lee, J.-B. Park, I.-K. Yoo, 2-stack 1D-1R cross-point structure with oxide diodes as switch elements for high density resistance RAM applications. IEDM Tech. Dig. 771–774 (2007)
51. M.-J. Lee, Y. Park, D.-S. Suh, E.-H. Lee, S. Seo, D.-C. Kim, R. Jung, B.-S. Kang, S.-E. Ahn, C.B. Lee, D.H. Seo, Y.-K. Cha, Two series oxide resistors applicable to high speed and high density nonvolatile memory. Adv. Mater. **19**, 3919–3923 (2007)
52. B.S. Kang, S.-E. Ahn, M.-J. Lee, G. Stefanovich, K.H. Kim, W.X. Xianyu, C.B. Lee, Y. Park, I.G. Baek, B.H. Park, High current-density $CuO_x/InZnO_x$ thin-film diodes for cross-point memory applications. Adv. Mater. **20**, 3066–3069 (2008)
53. G. Tallarida, N. Huby, B. Kutrzeba-Kotowska, S. Spiga, M. Arcari, G. Csaba, P. Lugli, A. Redaelli, R. Bez, Low temperature rectifying junctions for crossbar non-volatile memory devices, in IEEE International Memory Workshop Proceedings, pp. 6–8 (2009)
54. D. Kau, S. Tang, I.V. Karpov, R. Dodge, B. Klehn, J.A. Kalb, J. Strand, A. Diaz, N. Leung, J. Wu, S. Lee, T. Langtry, K.-W. Chang, C. Papagianni, J. Lee, J. Hirst, S. Erra, E. Flores, N. Righos, H. Castro, G. Spadini, A stackable cross point phase change memory. IEDM Tech. Dig. 617–620 (2009)
55. K. Gopalakrishnan, R.S. Shenoy, C.T. Rettner, K. Virwani, D.S. Bethune, R.M. Shelby, G.W. Burr, A. Kellock, R.S. King, K. Nguyen, A.N. Bowers, M. Jurich, B. Jackson, A.M. Friz, T. Topuria, P.M. Rice, B.N. Kurdi, Highly scalable novel access device based on mixed ionic electronic conduction (MIEC) materials for high density phase change memory (PCM) arrays. Symp. VLSI Tech. Dig. 205–206 (2010)
56. R.S. Shenoy, K. Gopalakrishnan, B. Jackson, K. Virwani, G.W. Burr, C.T. Rettner, A. Padilla, D.S. Bethune, R.M. Shelby, A.J. Kellock, M. Breitwisch, E.A. Joseph, R. Dasaka, R.S. King, K. Nguyen, A.N. Bowers, M. Jurich, A.M. Friz, T. Topuria, P.M. Rice, B.N. Kurdi, Endurance and scaling trends of novel access-devices for multi-layer crosspoint-memory based on mixed-ionic-electronic-conduction (MIEC) materials. Symp. VLSI Tech. Dig. 94–95 (2011)
57. M. Son, J. Lee, J. Park, J. Shin, G. Choi, S. Jung, W. Lee, S. Kim, S. Park, H. Hwang, Excellent selector characteristics of nanoscale VO_2 for high-density bipolar ReRAM applications. IEEE Electron Device Lett. **32**, 1579–1581 (2011)

58. D. Ielmini, S. Spiga, F. Nardi, C. Cagli, A. Lamperti, E. Cianci, M. Fanciulli, Scaling analysis of submicrometer nickel-oxide-based resistive switching memory devices. J. Appl. Phys. **109**, 034406 (2011)
59. R. Yasuhara, K. Fujiwara, K. Horiba, H. Kumigashira, M. Kotsugi, M. Oshima, H. Takagi, Inhomogeneous chemical states in resistance-switching devices with a planar-type Pt/CuO/Pt structure. Appl. Phys. Lett. **95**, 012110 (2009)
60. C.H. Kim, H.B. Moon, S.S. Min, Y.H. Jang, J.H. Cho, Nanoscale formation mechanism of conducting filaments in NiO thin films. Solid State Commun. **149**, 1611–1615 (2009)
61. D. Ielmini, Modeling the universal set/reset characteristics of bipolar RRAM by field- and temperature-driven filament growth. IEEE Trans. Electron Devices **58**, 4309–4317 (2011)
62. U. Russo, D. Ielmini, C. Cagli, A.L. Lacaita, Filament conduction and reset mechanism in NiO-based resistive-switching memory (RRAM) devices. IEEE Trans. Electron Devices **56**, 186–192 (2009)
63. C. Cagli, F. Nardi, D. Ielmini, Modeling of set/reset operations in NiO-based resistive-switching memory (RRAM) devices. IEEE Trans. Electron Devices **56**, 1712–1720 (2009)
64. Y.H. Tseng, C.-E. Huang, C.-H. Kuo, Y.-D. Chih, C.J. Lin, High density and ultra small cell size of contact ReRAM (CR-RAM) in 90nm CMOS logic technology and circuits. IEDM Tech. Dig. 109–112 (2009)
65. Y.H. Tseng, W.C. Shen, C.-E. Huang, C.J. Lin, Y.-C. King, Electron Trapping effect on the switching behavior of contact RRAM devices through random telegraph noise analysis. IEDM Tech. Dig. 636–639 (2010)
66. D.S. Jeong, H. Schroeder, R. Waser, Coexistence of bipolar and unipolar resistive switching behaviors in a Pt/TiO$_2$/Pt stack. Electrochem. Solid State Lett. **10**, G51–G53 (2007)
67. H. Schroeder, D.S. Jeong, Resistive switching in a Pt/TiO$_2$/Pt thin film stack—a candidate for a non-volatile ReRAM. Microelectron. Eng. **84**, 1982–1985 (2007)
68. L. Goux, J.G. Lisoni, M. Jurczak, D.J. Wouters, L. Courtade, C. Muller, Coexistance of the bipolar and unipolar resistive switching modes in NiO cells made by thermal oxidation of Ni layers. J. Appl. Phys. **107**, 024412 (2010)
69. L. Goux, Y. Chen, L. Pantisano, X. Wang, G. Groeseneken, M. Jurczak, D.J. Wouters, On the gradual unipolar and bipolar resistive switching of TiN\HfO$_2$\Pt memory systems. Electrochem. Solid State Lett. **13**, G54–G56 (2010)
70. D.C. Gilmer, G. Bersuker, H.-Y. Park, C. Park, B. Butcher, W. Wang, P.D. Kirsch, R. Jammy, Effects of RRAM stack configuration on forming voltage and current overshoot, in IEEE International Memory Workshop Proceedings, pp. 123–126 (2011)
71. J. Park, S. Jung, J. Lee, W. Lee, S. Kim, J. Shin, H. Hwang, Resistive switching characteristics of ultra-thin TiO$_x$. Microelectron. Eng. **88**, 1136–1139 (2011)
72. J. Lee, J. Shin, D. Lee, W. Lee, S. Jung, M. Jo, J. Park, K.P. Biju, S. Kim, S. Park, H. Hwang, Diode-less nano-scale ZrO$_x$/HfO$_x$ RRAM device with excellent switching uniformity and reliability for high-density cross-point memory applications. IEDM Tech. Dig. 442–445 (2010)
73. L. Goux, J.G. Lisoni, X.P. Wang, M. Jurczak, D.J. Wouters, Optimized Ni oxidation in 80-nm contact holes for integration of forming-free and low-power Ni/NiO/Ni memory cells. IEEE Trans. Electron Devices **56**, 2363–2368 (2009)
74. D. Ielmini, Filamentary-switching model in RRAM for time, energy and scaling projections. IEDM Tech. Dig. 409–412 (2011)
75. U. Russo, D. Kalamanathan, D. Ielmini, A.L. Lacaita, M. Kozicki, Study of multilevel programming in programmable metallization cell (PMC) memory. IEEE Trans. Electron Devices **56**, 1040–1047 (2009)
76. U. Russo, D. Ielmini, C. Cagli, A.L. Lacaita, Self-accelerated thermal dissolution model for reset programming in NiO-based resistive switching memory (RRAM) devices. IEEE Trans. Electron Devices **56**, 193–200 (2009)
77. D. Ielmini, C. Cagli, F. Nardi, Physical models of size-dependent nanofilament formation and rupture in NiO resistive switching memories. Nanotechnology **22**, 254022 (2011)

78. V.V. Zhirnov, R.K. Cavin III, S. Menzel, E. Linn, S. Schmelzer, D. Brauhaus, C. Schindler, R. Waser, Memory devices: Energy-space-time tradeoffs. Proc. IEEE **98**, 2185–2200 (2010)
79. S. Yu, H.-S.P. Wong, Compact modeling of conducting-bridge random-access memory (CBRAM). IEEE Trans. Electron Devices **58**, 1352–1360 (2011)
80. J. Park, M. Jo, E.M. Bourim, J. Yoon, D.J. Seong, J. Lee, W. Lee, H. Hwang, Investigation of state stability of low-resistance state in resistive memory. IEEE Electron Device Lett. **31**, 485–487 (2010)
81. D. Ielmini, F. Nardi, C. Cagli, A.L. Lacaita, Size-dependent retention time in NiO-based resistive switching memories. IEEE Electron Device Lett. **31**, 353–355 (2010)
82. D. Ielmini, Universal set/reset characteristics of metal-oxide resistance switching memories. ECS Trans. **35**, 581–596 (2011)
83. D. Ielmini, F. Nardi, C. Cagli, Resistance-dependent amplitude of random telegraph signal noise in resistive switching memories. Appl. Phys. Lett. **96**, 053503 (2010)

Chapter 7
Exploring Surfaces of Materials with Atomic Force Microscopy

S. Magonov, J. Alexander, and S. Belikov

Abstract Several aspects of Atomic Force Microscopy (AFM) are considered in this chapter. Theoretical backgrounds of AFM, which are based on the asymptotic solution of tip–sample interactions, lead to the classification of modes and computer simulations of images and force curves. Visualization of surface morphology with high resolution is the main AFM application. The practical issues of the high-resolution imaging, tracking of corrugated surfaces, and compositional mapping of multicomponent polymers are illustrated in several examples. The components of heterogeneous systems are recognized by their specific shape or by their different mechanical and electric properties revealed in the AFM-based methods. The challenges of the quantitative nanomechanical studies of soft materials are discussed. The multifrequency examination of local electric/dielectric properties is presented by the single-pass studies of surface potential and dielectric response on various samples.

7.1 Introduction

Microscopic studies of surfaces were revolutionized over 30 years ago with the invention of the scanning tunneling microscope (STM) [1]. In this device the detection of tunneling current between a sharp metallic tip and a (semi)conducting surface is applied for visualization of surface structures with atomic-scale resolution. Although the novel microscope has the same basic components of the earlier introduced topographiner [2], the innovative STM applications have advanced microscopic studies tremendously. At the beginning, the high-resolution visualization of surface structures with STM was restricted due to the need for sample conductivity. This limitation was removed when Atomic Force Microscopy (AFM) was introduced for

S. Magonov (✉) • J. Alexander • S. Belikov
NT-MDT Development Inc., 430 W. Warner Rd., Tempe, AZ 85284, USA
e-mail: magonov@ntmdt.us

A. Korkin and D.J. Lockwood (eds.), *Nanoscale Applications for Information and Energy Systems*, Nanostructure Science and Technology,
DOI 10.1007/978-1-4614-5016-0_7, © Springer Science+Business Media New York 2013

studies of surfaces with the use of more universal tip–sample forces [3]. In this technique, a micro-fabricated probe consisting of a cantilever with a sharp pyramidal tip at its free end serves as a sensor of forces acting between the tip apex (3–20 nm in diameter) and a sample surface. The tip–sample interactions are measured by the deflection of the cantilever that is monitored by an optical lever scheme [4]. At present, there are a number of different AFM scanning modes, in which the probe stays either in permanent contact with a sample or interacts intermittently, being driven into oscillatory motion. The changes of the cantilever deflection and oscillatory parameters (amplitude, frequency, phase) can be employed for monitoring tip–sample force interactions. For surface profiling, the probe or the sample is moved laterally, and the feedback adjusts their vertical separation to keep the force level at the set-point value. The map of the vertical adjustments represents the sample topography. Visualization of surface structures with high spatial resolution is achieved because the piezoelectric scanners with an accuracy better than 0.1 nm are used as the lateral and vertical actuators. The sensitivity of force measurements with probes, whose stiffness varies in the 0.03–40 N/m range, has reached the femto-Newton scale. AFM studies of surface structure and morphology are conducted for a broad range of materials, and such measurements can be performed in various environments (air, vapors, vacuum, liquid, etc.). The latter capabilities make this method unique among the microscopic techniques.

High-resolution and low-force profiling is the main but not the only function of AFM. Since the first applications it was found that the micro-fabricated probe with an appropriate coating can sense not only the mechanical tip–sample interactions but also the electric and magnetic forces. Therefore, the probe responds differently on surface locations with dissimilar properties, and this provides the AFM the ability to examine heterogeneous samples and distinguish their components. The compositional mapping of multicomponent materials became the important function often used in the examination of morphology and composition of industrial samples. Such studies are crucial for finding the structure–property relationship. The differentiation of individual components of complex materials can be achieved by measuring local changes of mechanical properties with the force modulation mode or phase imaging as well as by detecting the variations of surface potential and dielectric response. In some cases, the placement of a polymer sample into a vapor of organic solvent leads to selective swelling of a particular component that will alternate its mechanical and electric property and the change can be detected with AFM. The sensitivity of the method to local properties raises a question about their quantitative measurements. This subject becomes of increasing importance in a number of industries as the dimensions of functional structures continuously shrink. The challenges of quantitative mechanical and electric measurements with AFM will be discussed below.

The goal of this chapter is to overview AFM capabilities in the visualization of surface morphology, high-resolution imaging, compositional mapping, and quantitative studies of the local materials properties. Before the analysis of these topics, we will present the mathematical description of the AFM probe behavior and its interactions with a sample. The practical issues of the method will be covered in the part related to studies of surface morphology and high-resolution imaging. These issues include the choice of AFM mode, the applied probe, and optimization of the experimental parameters. The other two topics are related to AFM-based studies of mechanical and electric properties of materials. In studies of local mechanical properties of soft materials, the experiment should be accompanied by a thorough characterization of the probe stiffness and shape as well as by a rational choice of the deformation model for data analysis in terms of elastic modulus and work of adhesion. In the description of AFM-based electric modes, we will focus on the measurements of electrostatic tip–sample force and force gradient for measurements of surface potential and local dielectric response. For independent and simultaneous detection of the mechanical and electrostatic forces we applied the multifrequency approach, which becomes a powerful tool for advancing AFM studies. This method allows single-pass measurements of sample topography, surface potential, and dielectric response. All topics of AFM studies will be illustrated by practical examples. In the conclusion, we summarize the current state of AFM and related methods and outline further development trends in their instrumentation and applications.

7.2 Theoretical Backgrounds and Modeling of Atomic Force Microscopy

AFM was introduced as a contact mode technique and its value has been demonstrated in numerous applications. With developments of different modes the method becomes more informative and more complicated. The use of an oscillating probe in AFM has expanded its applicability but it has a serious drawback due to the complex relationship between the parameters of probe dynamics (amplitude, phase, frequency) and the tip–sample force. For a while, a poor knowledge about the tip–sample force interactions in oscillatory modes was compromised by the empirical observations and simplified theoretical considerations [5–7]. Although all of these efforts were very useful and informative, a rigorous approach towards dynamics of AFM modes is still required. The theoretical understanding of the behavior of the probe interacting with a sample becomes pivotal for quantitative measurements of the tip–sample forces of different origin (mechanical, electrostatic, magnetic). In many occasions, these forces are acting simultaneously and figuring out their contributions is a challenging task. In this part we describe a model of AFM, in which the probe behavior is treated in terms of asymptotic nonlinear mechanics. The description of the model will be followed up by the classification of oscillatory modes. Afterwards we will consider different types of tip–sample interactions and their analysis. The computer simulation of the AFM operation in different modes and its applications to the force curves and surface imaging will be presented at the end. The use of the theoretical model in the interplay with the experimental data

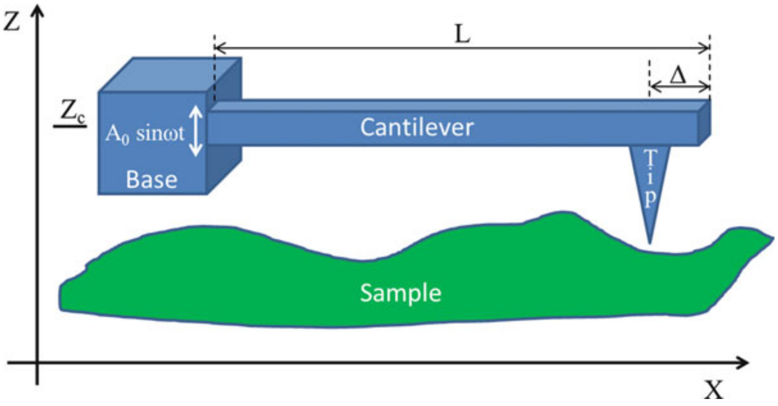


Fig. 7.1 A sketch of the atomic force microscopy probe near a sample surface

provides a background for better understanding of AFM functionality, realistic simulators, improved control design, and quantitative measurements of local properties.

7.2.1 Basic Model of AFM

A mechanical model of the AFM probe that will be considered in the static and dynamic analysis is presented in Fig. 7.1. A thin cantilever of length L and cross-section $S = b \times h$ is attached to the base at point $(x = 0, Z = Z_c)$, where Z_c is the cantilever position at rest. In the contact mode the cantilever rests before the tip touches a sample and in oscillatory modes the base is oscillating as $Z(0,t) = Z_c + A_b \sin \omega t$, where A_b and ω are the amplitude and frequency of the piezo-driver shaking the probe. Concentrated forces may be applied to the tip due to tip–sample interactions at the distance $\varDelta$ from the edge, i.e., at $x = L - \varDelta$ (L—the probe length); the weight of the tip is applied at the same point.

Mathematically this problem can be described by an Euler–Bernoulli's type of equation [8] that is shown below.

$$\frac{\partial^2 Z(x,t)}{\partial t^2} + 2\beta \frac{\partial Z(x,t)}{\partial t} + a^2 \frac{\partial^4 Z(x,t)}{\partial x^4} = [H(Z(x,t)) - p]\delta(x - (L - \Delta)). \quad (7.1)$$

In case of the contact mode (quasi-static situation) the dependence of Z on time can be neglected. In the oscillatory mode we need to consider the following boundary conditions:

$$\text{Base motion:} \qquad Z(0,t) = Z_c + A_0 \sin \omega t. \quad (7.2)$$

$$\text{Cantilever is attached to the base:} \qquad \frac{\partial Z}{\partial x}(0,t) = 0 \tag{7.3}$$

$$\text{Free end (no moment):} \qquad \frac{\partial^2 Z}{\partial x^2}(L,t) = 0 \tag{7.4}$$

$$\text{Free end (no force):} \qquad \frac{\partial^3 Z}{\partial x^3}(L,t) = 0 \tag{7.5}$$

The following notations (SI units are shown in brackets) are used in Eqs. (7.1)–(7.5): b and h [m] are the width and height of the cantilever; x [m] the horizontal coordinate; Z [m] the vertical coordinate; t [s]-time; $Z(x,t)$ [m] the vertical position of the cantilever at point x and time t; $\beta[\mathrm{s}^{-1}]$ the damping coefficient (may depend on ω); $a^2 = \frac{EI}{\rho S}\left[\frac{m^4}{\mathrm{s}^2}\right]$, where $E\left[\frac{N}{m^2}\right]$ is Young's modulus of the cantilever; $I = \frac{bh^3}{12}\,[\mathrm{m}^4]$ the moment of inertia of the cross-section with respect to horizontal axis; $\rho\left[\frac{kg}{m^3}\right]$ the density of the cantilever; $S = bh[\mathrm{m}^2]$ the area of the cross-section of the cantilever; $H(Z) \times \rho \times S[N]$ the concentrated force applied to the tip due to tip–sample interaction; $p \times \rho \times S\,[N]$ the weight of the tip; and $\delta(x)\ \left[\frac{1}{\mathrm{m}}\right]$ the Dirac's delta function.

Relevant to the contact mode of operation, the solution of the stationary Euler–Bernoulli equation gives the dependence of the cantilever deflection on the force with the coefficient known as the inverse of the spring constant, k_{static}^{-1}, where the spring constant is

$$k_{\mathrm{static}} = \frac{bh^3E}{4L^3}. \tag{7.6}$$

The case of oscillatory modes is more complicated. Ignoring the weight of the tip (which adds only an insignificant technical complication), the motion of the tip located at the position $l = L - \Delta$ near the first resonance is

$$Z(l,t) = Z_{\mathrm{c}} + \eta_1 + \xi_1 \tag{7.7}$$

where ξ_1 and η_1 satisfy the following equations

$$\ddot{\xi}_1 + 2\beta_1\dot{\xi}_1 + \omega_1^2\xi_1 = \frac{F\Big(Z_{\mathrm{c}} + \eta_1 + \xi_1;\ \mathrm{sgn}\Big(\dot{\eta}_1 + \dot{\xi}_1\Big)\Big)}{\rho S\left\|Z_1^l\right\|^2} \tag{7.8}$$

$$\eta_1 = A_0\cos(\omega t + \varphi_0) \tag{7.9}$$

and ω_1 is the 1st Eigen-frequency of the cantilever [8];

β_1, calculated by the formula

$$\beta_1 = \frac{\omega_1}{\sqrt{4Q_1^2 - 1}} \approx \frac{\omega_1}{2Q_1} \tag{7.10}$$

is the damping factor of the cantilever's 1st mode and Q_1 is the quality factor of the 1st mode; $\left\|Z_1^l\right\|$ is a norm of the special basis function of the first mode such that $Z_1^l(l) = 1$; the norm has SI unit $[\sqrt{\mathrm{m}}]$; A_0 and φ_0 are (calculated using Euler–Bernoulli based mathematics) the amplitude and phase of the tip's oscillation caused by the base oscillation only (usually called "free" oscillation by AFM practitioners), but not by tip–sample interactions; A_0 and φ_0 depend on A_0, ω, ω_1, β_1, and Δ; and $F(z;\operatorname{sgn}\dot{z})$ is the tip–sample interaction function at the cantilever position z at the location of the tip; this function defines the behavior of the mode; we assume that it depends on the direction of the motion $\operatorname{sgn}\dot{z}$.

The rational approximation of the vibrational equations was obtained in the framework of the asymptotic Krylov–Bogoliubov–Mitropolsky (KBM) method [9]. The main assumption $\varepsilon = Q_1^{-1} \ll 1$ is correct for AFM operations in ambient conditions and vacuum. It was also assumed that the difference between the driving frequency ω and the Eigen-frequency ω_1 is of the order ε and define the related parameter g of order 1 as

$$g = \frac{\omega - \omega_1}{\varepsilon} = Q_1(\omega - \omega_1)\,. \tag{7.11}$$

Steady-state solutions of asymptotic dynamics for the cantilever amplitude and phase (x,θ) satisfy the following system of equations:

$$\begin{cases} \sin\theta = \dfrac{1}{\sqrt{1+\left(\dfrac{2g}{\omega_1}\right)^2}}\left\{\frac{1}{N}\int_0^\pi [F_\mathrm{a} - F_\mathrm{r}](Z_\mathrm{b} + A\cos y)\sin y\,\mathrm{d}y + \frac{A}{A_0}\right\} \\ \cos\theta = -\dfrac{1}{\sqrt{1+\left(\dfrac{2g}{\omega_1}\right)^2}}\left\{\frac{1}{N}\int_0^\pi [F_\mathrm{a} + F_\mathrm{r}](Z_\mathrm{b} + A\cos y)\cos y\,\mathrm{d}y + \frac{2g}{\omega_1}\frac{A}{A_0}\right\}, \end{cases} \tag{7.12}$$

where F_a and F_r are the tip–sample forces acting during tip approach to the surface and its removal; A_0 is the amplitude of the cantilever before it is engaged into the tip–sample interactions.

7.2.2 *Classification of AFM Modes*

The equations thus obtained can be used for the classification of AFM modes. The two expressions forming Eq. (7.12) have four unknowns: amplitude A, phase θ, base position Z_c usually called height, and variable g proportional to frequency shift. The "free" oscillation amplitude A_0 is assumed constant, however it can be considered as a fifth unknown (in this case usually A assumed constant). To solve these equations, two of the four unknowns should be kept constant and this provides six possible combinations. Each combination relates to a potential AFM dynamic control mode and four of these modes are widely used. In practice two unknown variables are kept constant by feedback control. For simplification the feedback control is assumed instant and ideal. This allows avoiding dealing with the complications of feedback implementation and concentrating on the definition of the dynamic control modes and their properties.

The four widely used dynamic control modes are Amplitude Modulation (AM) and Frequency Modulation (FM) in their imaging and spectroscopy operations. Imaging in the AM mode is defined by g = constant (usually $g = 0$—resonance) and $A = A_{\text{set-point}}$ = constant. At each sample position XY the tip–sample interaction determines F_{a} and F_{r} in Eq. (7.12); solutions Z_b versus XY give height image and θ versus XY—phase image—the main data channels for AM. In AM spectroscopy mode (g and Z_b = const) for each Z_b the amplitude and phase values (A and θ) are calculated, respectively, to get the A versus Z_b (AvZ) and phase versus Z_b (θvZ) dependencies. In the Frequency Modulation (FM) mode, imaging is performed under the feedback keeping a constant phase (usually $\theta = \pi/2$—resonance) and g. For each XY point of the sample, tip–sample force determines F_{a} and F_{r} in Eq. (7.12); solutions Z_b versus XY give height image; and A versus XY –amplitude image—the main data channels for the FM mode. In practical applications at ambient conditions the AM mode is the most dominant mode. The FM mode is invaluable for studies in UHV where high quality factor makes the AM mode impractical. The use of the FM mode for experiments in air and under liquid [10] has been initiated in a number of laboratories yet it did not become the routine method.

As we see from the above equations in (7.12), AFM tip–sample interactions are very critical for interpretation of experimental results, and so it is a topic of intensive experimental and theoretical study. At this moment we will consider mostly Lennard–Jones solid and elastic solid models. The specific of AFM is in its operation at spatial dimensions where molecular level interactions converge with those considered for solid matter. The Lennard–Jones potential is a model of interaction on the molecular level. The attraction force per unit area between two parallel plates and related potential can be computed by integration ([11], p. 31). Then, using the Derjaguin approximation, the force between a parallel layer and the spherical tip of molecules can be modeled by the following formula ([11], p. 31)

$$F(z) = 2\pi R\frac{8\gamma}{3}\left[\frac{1}{4}\left(\frac{Z_0}{z}\right)^8 - \left(\frac{Z_0}{z}\right)^2\right] \tag{7.13}$$

where z is the distance between the layer and the tip and the parameters are: Z_0 is the equilibrium distance, γ is a half of the work of adhesion, and R is the radius of the tip. A realistic correction to model equation (7.13) for molecular multilayers such that $N/2$ layers are free to move has been described in ([11], pp. 39–45). The combination of Lennard–Jones, Derjaguin approximation, and adhesive avalanche describes the tip–sample interaction before and slightly after the point of geometrical contact [12]. Mechanical (elastic and elasto-adhesive) contact is described by elastic solid models (Hertz [11], p. 245), JKR ([11], p. 272, 269), and DMT ([11], p. 238). These models depend on radius of the tip R, work of adhesion $\omega = 2\gamma$, and reduced elastic modulus E^{*}. The JKR–DMT transition is parameterized by λ ([11], pp. 290–292). Some generalizations of these models for the tips with arbitrary axisymmetrical profiles are presented in [12, 13]. These elasto-adhesive models describe the tip–sample interaction after and slightly before the point of geometrical contact. It seems natural to match the Lennard–Jones solid curves with the elastic solid curves in the area slightly before and slightly after the point of geometrical contact where both models are valid. This match is demonstrated in [12] and can be implemented algorithmically on experimental curves by fitting parameters N and λ. It provides a consistent and theoretically justified model of tip–sample interactions from the tip approaching the surface to a deep penetration into the sample. The model reveals the qualitative features of the force curves but requires thorough experimental verification.

7.2.3 Examples of AFM Simulations

The theoretical background discussed earlier was implemented into a LabVIEW-based computer program, which can be applied for the simulation of AFM experiments. The interplay between the computer simulation and practical results has been realized for several cases. In these simulations, the Hertz model was employed for tip–sample force interactions. Figure 7.2a presents the AvZ and θvZ curves in the AM spectroscopy mode, which are simulated for a nondissipative case [14]. Different solutions (branches) for amplitude and phase dependencies on Z are named as the high amplitude branch H, low amplitude branch L, and unstable branch U. On a sample approach to the oscillating probe [the distance changes from right (large Z) to left (small Z)] the system accommodates at the L branch with amplitude and phase both gradually diminishing. As Z is changing a saddle–node bifurcation in amplitude/phase coordinates can happen and the system will jump to the H branch. The reverse transition is expected at lower Z. This behavior well matches the experimental curves obtained on a Si substrate, Fig. 7.2b. The locations with the abrupt transitions are marked with the circles. The most drastic changes are observed for the θvZ curves, and they are also related to the transition of the tip–sample forces from the attractive to repulsive range. When a dissipative case of adhesive avalanche [11] was considered, the main features of the simulated curves remained the same as in the nondissipative case yet their shape has changed, Fig. 7.2c. Experimental data obtained on polymer samples are consistent with these theoretical predictions although variations were observed depending on type of sample, stiffness of the probe, and tip size. In some cases such as operation at small

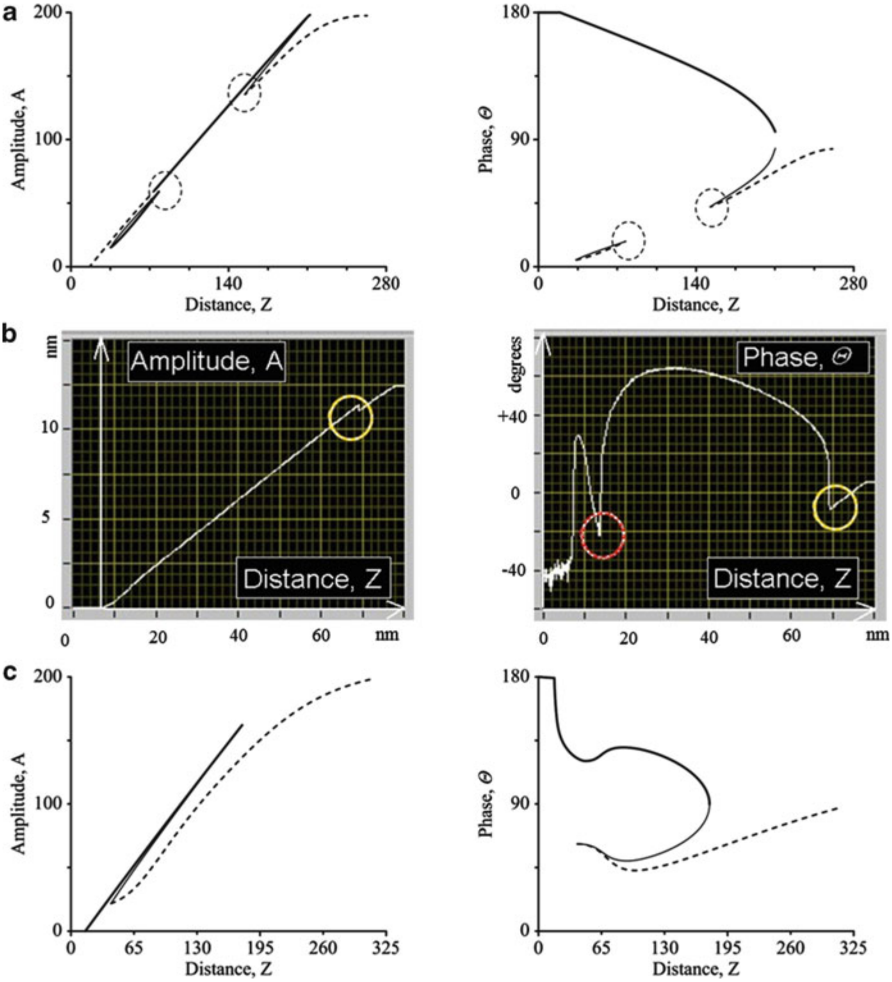


Fig. 7.2 Simulated (**a**) and experimental (**b**) *A*v*Z* and *P*v*Z* curves. Simulations were done for the conservative case. (**c**) Simulated *A*v*Z* and *P*v*Z* curves for the dissipative case of adhesive avalanche. In the simulated curves the high amplitude (H), low amplitude (L), and unstable (U) branches are indicated with *thick*, *dashed,* and *thin lines*, respectively. The *circles* indicate the regions where bifurcations take place. The measurements were performed on a Si substrate

amplitudes, and imaging of sticky materials, the amplitude and phase response behave according to the L branch solution. The abrupt changes of the phase curves and the kinks on the amplitude curves, which are consistent with the transition from the L branch to the H branch, were also often detected. These effects are seen as image artifacts and can be avoided by choosing set-point amplitudes away from the transition region. The U branch is unstable and so was never detected by experiment; it is important, however, for describing the bifurcations.

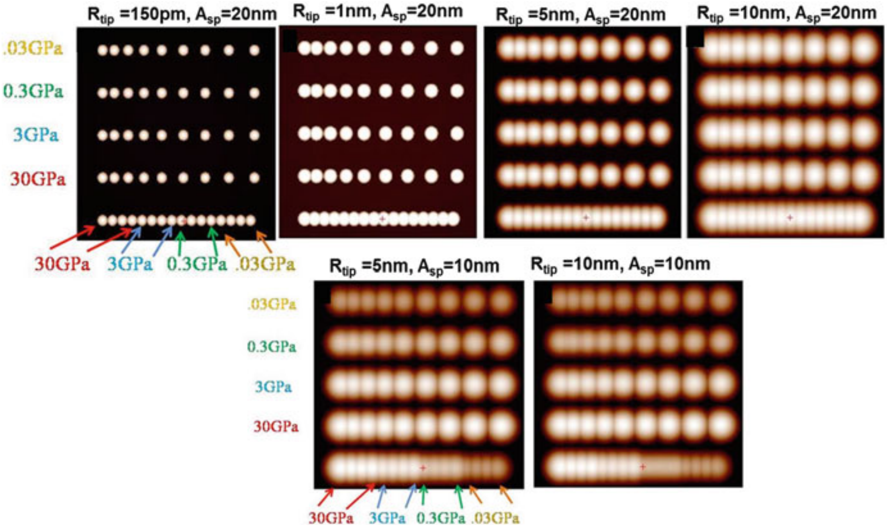


Fig. 7.3 (**a**–**e**) Computer simulation of AFM patterns of spherical objects with different stiffness and in various geometric arrangements. The calculations were performed for AM mode using Hertz model for tip–sample force interactions. The elastic modulus of the spheres and tip radius and set-point amplitudes are indicated on the images. The scan size is120 nm × 120 nm

Simulation of the images in different AFM modes might shed light on many outstanding problems of these techniques related to experiments at different forces and the use of probes of various geometries and tip dimensions. The effect of the tip dimensions and the imaging force is well distinguished in the AFM height patterns of individual spherical objects, which have the same dimensions but different arrangements and various elastic moduli, Fig. 7.3a. The stiffness of the spheres is chosen in the 0.03–30 GPa range (the rows, top to bottom), and they were placed at different spacing between each other. In the bottom row, the blocks of four spheres with different modulus are tightly packed. The AM-mode scanning of these spheres with an atomically sharp probe and at low tip-force ($R = 0.15$ nm and A_{sp} is close to A_0) gives the image with the correct dimensions of the objects, Fig. 7.3a. The images in Fig. 7.3a, b point out how the spheres' appearance is changing with the increase of the tip dimensions at minimal tip-force. The images in Fig. 7.3d, e show the effect of the tip-force. Such image simulations can be helpful for correct assignment of the objects' dimensions deduced from AFM images of various samples and for optimization of imaging conditions.

Another example of image simulation is related to the observations of atomic-scale or molecular-scale defects in AFM images. A recording of such defects in the images of the periodic lattices is often considered to be the proof of true atomic- or molecular-scale resolution. We modeled the AM images of the polydiacetylene (PDA) crystal surface, in which two molecules were removed from different sample locations [15]. The simulations were performed for tips with radii of 150 pm and 1 nm, and for a tip with a radius of 5 nm for operation at low ($A_{sp} \sim A_0 = 20$ nm) and slightly elevated ($A_{sp} = 19$ nm, $A_0 = 20$ nm) forces. The results shown in

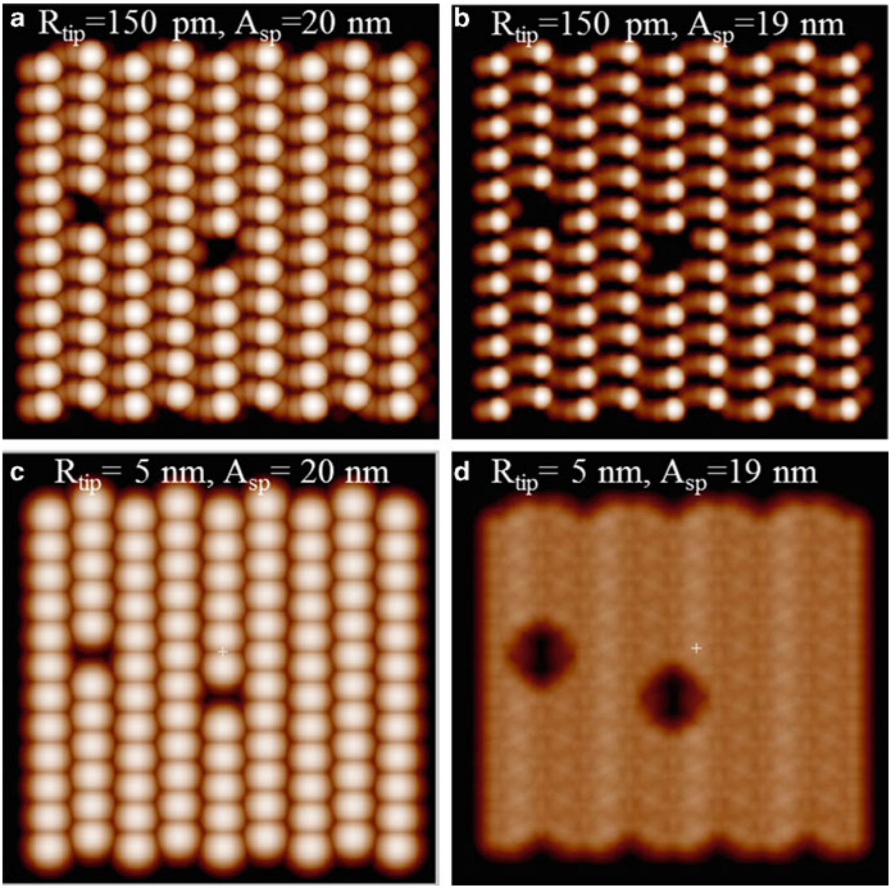

a R_{tip}=150 pm, A_{sp}=20 nm
b R_{tip}=150 pm, A_{sp}=19 nm
c R_{tip}= 5 nm, A_{sp}= 20 nm
d R_{tip}= 5 nm, A_{sp}=19 nm

Fig. 7.4 Computer simulated images of the *bc* plane of polydiacetylene crystal. Two defects related to missing molecules were introduced into this lattice. The images in (**a**, **b**) were simulated for the atomically sharp tip ($R_{\text{tip}} = 150$ pm) acting at low and slightly elevated forces ($A_{\text{sp}} = 20$ nm and $A_{\text{sp}} = 19$ nm). The images in (**c**, **d**) were simulated for the tip with radius $R_{\text{tip}} = 5$ nm acting at low and slightly elevated forces ($A_{\text{sp}} = 20$ nm and $A_{\text{sp}} = 19$ nm)

Fig. 7.4a–d indicate that single molecular defects are clearly distinguished in the patterns generated for all cases. The defects are seen as spots of 0.25, 0.1, and 0.02 nm in size for tips with $R = 150$ pm, 1 nm, and 5 nm, respectively. This finding suggests that the visualization of single-molecule defects is quite possible even with tips of 5 nm in radius, yet the instrumental noise and thermal drift may introduce practical difficulties. It is worth noting that the lattice pattern around the missing molecules correctly reproduces the surface molecular structure only in the images shown in Fig. 7.4a, b. In other cases (Fig. 7.4c, d) the fine structure of the pattern changes most likely due to bifurcations [15]. This leads us to the conclusion that the observations of such defects do not guarantee that the pattern surrounding the defects correctly reflects the surface molecular structure.

7.3 Studies of Surface Morphology and High-Resolution Imaging

7.3.1 *Optimization of the Experiment*

The interest of atomic-scale imaging was the main motivation behind the development of AFM. In practice, the method becomes even more useful for surface microscopy by providing the complementary information to surface profilers and other microscopic techniques. Initially AFM was used as a contact mode technique, in which the tip-force in the vertical direction is used for the surface-tracking feedback and lateral force variations are monitored to reflect the tip–sample friction. Unfortunately, shearing forces can be destructive for soft materials and this limitation is practically absent in the oscillatory AM and FM modes. Therefore, the AM mode, which was first known as the tapping mode [16], is applied in the majority of applications. Typically, AFM modes are applied for scanning samples over areas up to 100 µm on a side, which corresponds to the largest extensions of the lateral scanners. In profiling of surface corrugations, the limitations are related to the range of the vertical piezo-element (up to 15 µm) and the length and shape of the tip. Most commercial probes have tips with the height in the 7–10 µm range, and the tip opening angle can be as small as 10°. The tip geometry in commercial micro-fabricated probes made of Si_3N_4 and Si crystals is guided by the crystalline structure of these materials. Therefore, the opening angles in these probes are varied from 90° to 35°. Probes having tips with a higher aspect ratio are specially made by focused ion beam (FIB) etching and diamond polishing [17]. For high-resolution imaging, which is achieved on atomically smooth surfaces, the size of the tip apex is the main factor. Most sharp tips have an apex of 3–5 nm in size. The proper choice of the probe is only one of the factors in optimization of AFM experiments. Adjustment of the tip–sample force is another major experimental issue. Minimization of the tip-force is essential already done in its approach to the sample. A gentle tip engagement procedure, which in the AM mode can be based on a monitoring the phase of the oscillating probe approaching the surface, allows avoiding tip breakage in the first contact with the sample. During scanning a minimal tip-force limits an undesirable disturbance of soft materials and reduces a tip–sample contact area, thus improving the spatial resolution. In AFM compositional imaging of heterogeneous materials, the tip-force increase will enhance the variations of local mechanical responses of the individual components of these materials. Therefore, imaging at different tip-forces particularly for soft materials brings valuable information. Another experimental factor to consider is the scanning rate, which is related to the feedback mechanism and imaging conditions. In the AM mode when the probe is moved to a new location its amplitude is measured and compared to the set-point value. For a probe oscillating at its resonance it takes some time (depending on its quality factor) to adopt a new value and this is the limiting factor to fast scanning especially on corrugated surfaces. Practically, a comparison of trace and retrace profiles is a good criterion of the profiling precision and in the AM mode the appropriate scanning rates are in the 0.5–2.0 Hz range.

In many cases, AFM imaging of samples does not require any preparative efforts. Although any kind of material can be examined with this technique, for some samples, however, sample preparation might be necessary. For making samples best suited for AFM imaging, single macromolecules, organic and polymer thin films are spin cast from their dilute solutions in different solvents on flat substrates, such as mica, graphite, Si wafer, etc. Relatively smooth polymer surfaces for AFM visualization can be also prepared by hot melting of polymer semicrystalline samples between two flat substrates then followed by their cooling to room temperature. Elastomers and materials with rubbery components require more preparative efforts with the use of a cryo-ultramicrotome equipped with a diamond knife. A low-temperature (below glass transition of the rubber components) cutting of such materials provides smooth surfaces for imaging. Studies of such samples should be done immediately after the preparation otherwise low surface energy material exudes to the surface and makes its morphology different from that in bulk.

7.3.2 Examining Corrugated Samples

The previously discussed aspects of AFM imaging are illustrated below using results obtained on samples with morphology of different complexity and on crystalline structures and individual macromolecules, where high-resolution data can be obtained. First, we consider the AFM height images obtained on various microporous membranes, which are shown in Fig. 7.5a, b. The surface corrugations of the nitrocellulose membranes, which are revealed in Fig. 7.5c, are relatively high, reaching several microns in Fig. 7.5a and half of a micron in Fig. 7.5b. The profiling of these pores demands the use of tips with a high-aspect ratio. The above images were obtained with diamond tips having the opening angle of 9°. When the same samples were examined with regular Si probes whose tip has an opening angle of ~35°, then the image artifacts showing convolution of the tip shape with the surface features dominated the images, Fig. 7.5c.

The image of polypropylene microporous membrane Celgard™ 2400 shows the surface corrugations between fibrillar and lamellar regions, which are in the 20–30 nm range, Fig. 7.6a. This material is much smoother and the regular Si probes can be applied for visualization of the membrane morphology. Tip-force control is more important for this sample because the sharp tip that penetrates into the inter-fibrillar regions during scanning might push the fibrils aside thus making cavities that do not exist in reality. The limitations of the tip geometry can be noticed not only on surfaces with large corrugations but also on samples where

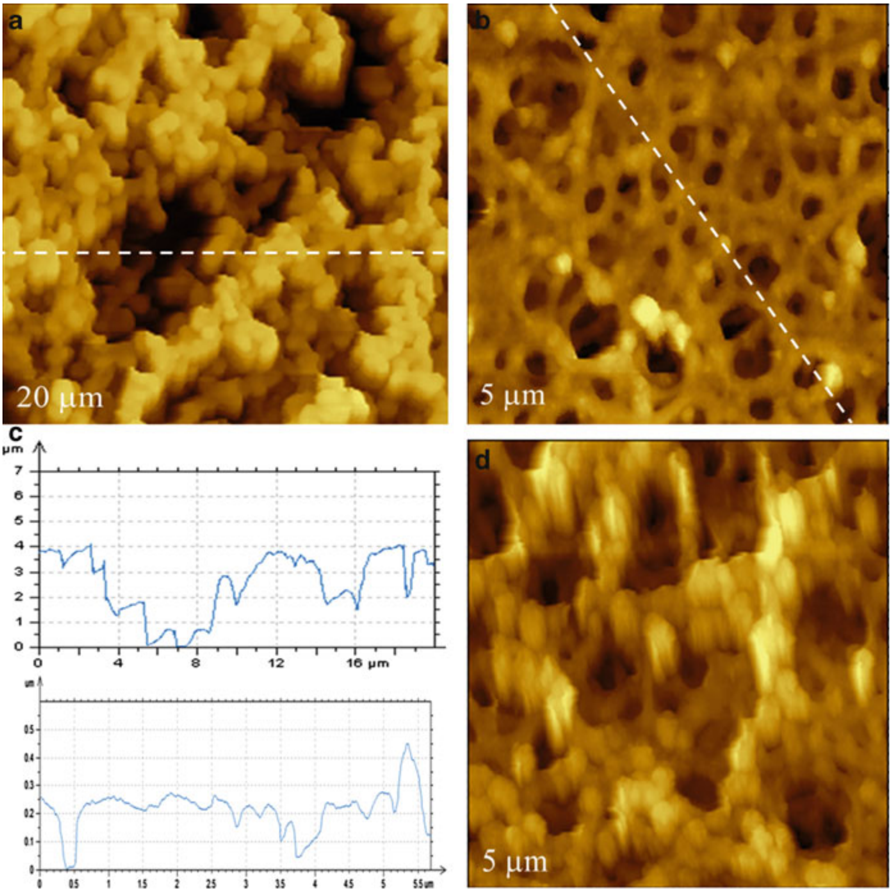


Fig. 7.5 (**a**, **b**) AFM height images of nitrocellulose membranes obtained with the tip having the opening angle of 9°. The cross-section profiles taken in these images across the directions indicated with white dashed lines are presented in (**c**). The height image in (**d**) was obtained with a regular Si probe that has a tip opening angle of 30°. The size of the image side is indicated in the *bottom left corner* of the images

crystalline structures have steep edges at the nanometer scale. An example of such morphology is presented in Fig. 7.6b with the image of SiGe crystalline features from epitaxially grown crystals on a Si substrate. The surface corrugations of this sample, which cover the range of several tens of nanometers, are relatively steep. These features were more correctly resolved in this image, which was recorded with the high-aspect tip. The comparative analysis of this image with another one obtained with the regular Si tip revealed that the volume of surface wells is 10% higher in the pattern obtained with the sharper probe.

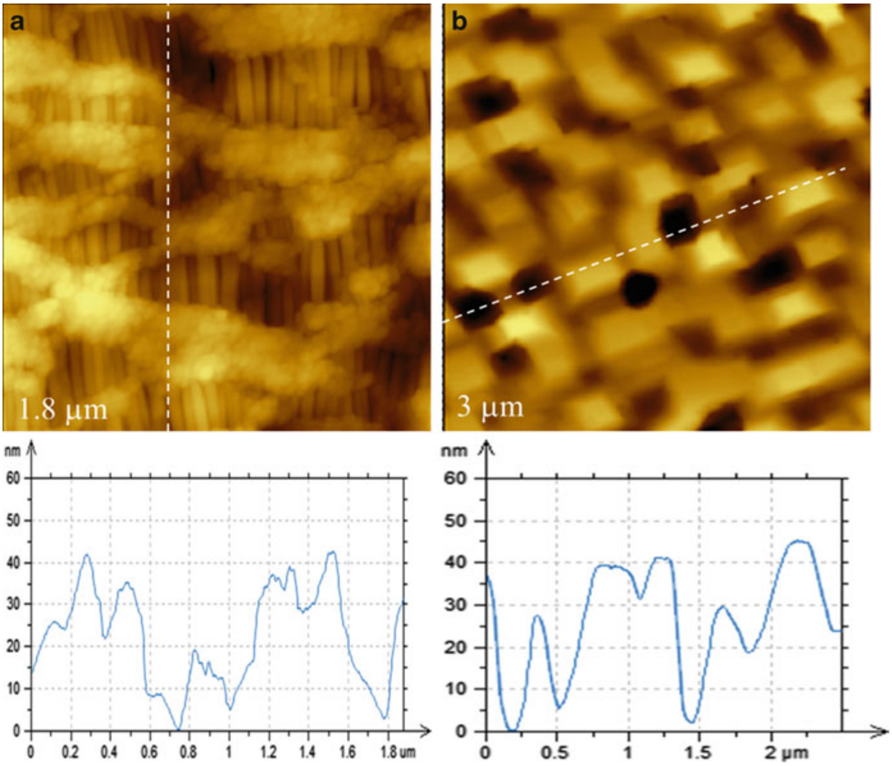


Fig. 7.6 Height images of Celgard 2400™ microporous membrane (**a**) and SiC crystals epitaxially grown on Si substrate (**b**). The cross-section profiles taken in these images across the directions indicated with *white dashed lines* are presented underneath the images

7.3.3 Morphology of Polymer Materials

The images of semicrystalline polymers poly(vinyledene fluoride) PVDF and high-density polyethylene (HDPE) are presented in Fig. 7.7a–d. They illustrate the morphology features and nanostructures that are typically recorded with AFM on semicrystalline polymer samples. Polymer spherulites are common for crystallization of macromolecules, and they are formed of lamellar structures in which polymer chains are multiple folded. The large-scale height image reveals one of the PVDF spherulites Fig. 7.7a. The cross-section across the spherulite (Fig. 7.7b) shows the height variations in the micron range, yet the local corrugations are not pronounced. The lamellar structure of the PVDF spherulite is magnified in the 2-µm image, which is taken on one of the "beams" running from the spherulite's center, Fig. 7.7c. The fine structure is quite diverse and, in addition to multiple fibers, one can distinguish the lamellar sheets that are overlaying each other in the location surrounded by a white dashed circle. The individual edge-on lamellae, which are

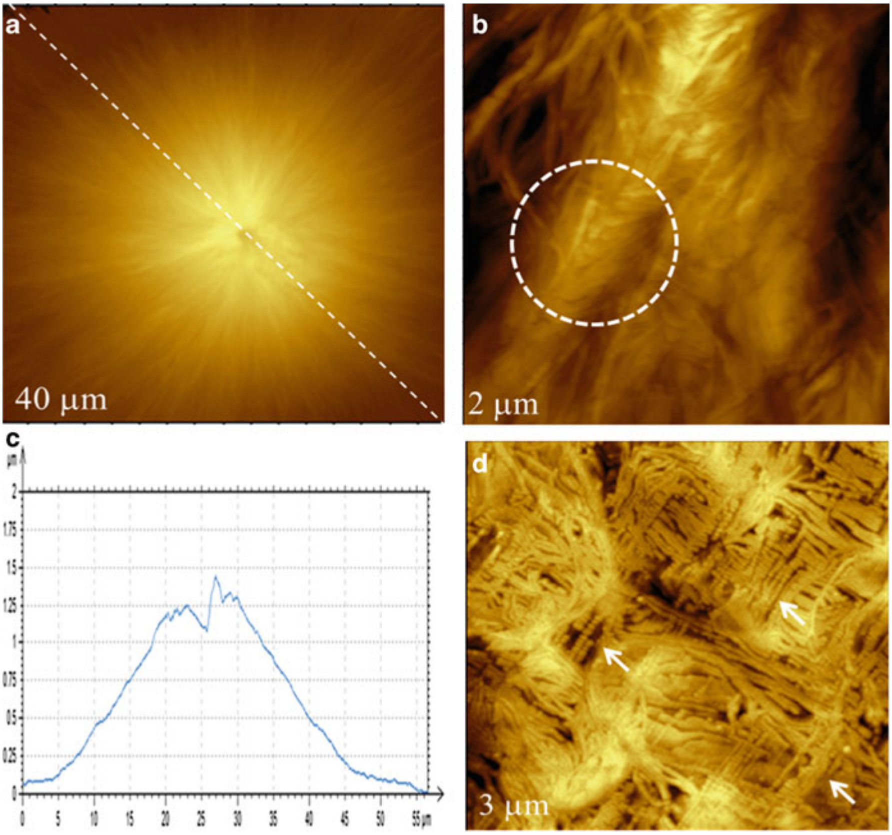


Fig. 7.7 Height images of a poly(vinyledene fluoride) crystal in (**a**) and (**c**). The cross-section profile along the diagonal direction in *A*, which is marked by a *white dashed line*, is shown in (**b**). A location with the array of flat lying lamellae is marked by a *white dashed circle* in (**c**). Height image of the lamellar and shish-kebab structures of HDPE is in (**d**). The *white arrows* point out the shish-kebab structures

15–25 nm in width, are clearly distinguished in this image. Such individual lamellae might be lying flat or edge-on, and the regions with periodically changing lamellar orientation are common to banded spherulites [16]. The lamellar structure of the polymer is directly related to crystallization conditions and varies from polymer to polymer. Among the crystalline structures of HDPE, which was crystallized by fast cooling from the melt, one can see not only lamellar blocks forming wavy patterns but also fibrils and "shish-kebab" sequences, Fig. 7.7d. The latter, which are pointed out in the image by several white arrows, are formed by crystallization of extended polymer chains into the fibrils ("shish") and coiled polymer chains into the chain-folded blocks (the short lamellae or "kebab").

As we learned from above, the polymer lamellae are recognized in AFM images based on their shape and characteristic dimensions, which are commonly

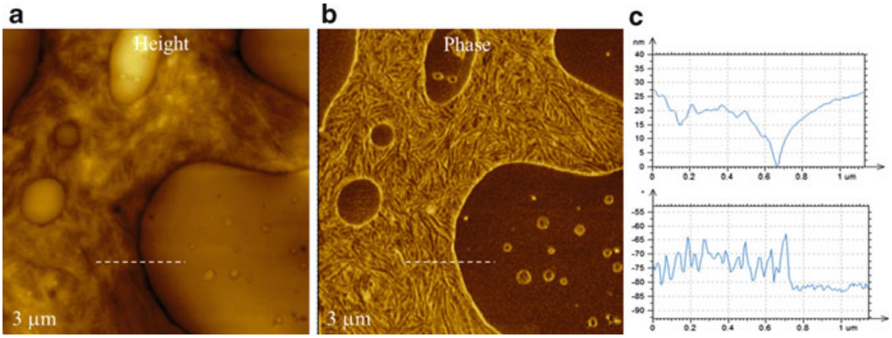


Fig. 7.8 Height and phase images (**a**, **b**) obtained at elevated-force imaging of a sample of PS/LLDPE blend. The cross-sections taken across the directions shown by the *white arrows* are presented in (**c**)

determined from other microscopic and diffraction techniques. The crystallinity of polymers is not very high, and amorphous components of different polymers are characterized by glass transitions (T_g) below and above room temperature (RT). This circumstance should be taken into account for the optimization of AFM imaging of semicrystalline polymers. In these materials lamellae are surrounded by amorphous polymer, which is substantially softer when in its rubbery state. Therefore, the visualization of lamellar structures in polymers with a T_g below RT (e.g., polyethylene, polypropylene) is facilitated by the fact that the nearby amorphous material is depressed by the tip especially at elevated forces. Indeed, the contrast differentiating the lamellar structures of these polymers in the height and phase images can be improved with a tip-force increase. Although the tip-force in oscillatory mode varies in each cycle, the study of a model sample made of layers of polyethylene with different density showed that an averaged tip-force level is minimal at small amplitude damping ($A_{sp} = 0.8–0.9\ A_0$) and low A_0 [18]. It was also shown that the tip–sample forces are most elevated when $A_{sp} = 0.4–0.5\ A_0$. The force level also depends on stiffness of the applied probe. However, the use of soft probes in ambient conditions is often restricted by a sample adhesion and capillary effects that prevent a stable probe oscillation. The phase images, which reflect the changes in phase of the oscillating probe when it comes into strong force interactions with a sample, are very sensitive to variations of local mechanical properties. Such images are widely used for compositional mapping of heterogeneous polymer materials. For polymers with a T_g higher than RT [e.g., poly (ethylene terephthalate), polystyrene—PS] the imaging at temperatures above the T_g provides high-contrast patterns revealing their lamellar organization. The compositional mapping with AFM is demonstrated by height and phase images of an immiscible blend of linear low density polyethylene (LLDPE) and PS, Fig. 7.8a, b. The images were obtained at elevated tip-force, and the phase contrast is substantially different from the corrugations revealed in the height image. The latter shows the elevated domains embedded into a matrix. A comparison of the cross-section profiles, which were taken along the direction marked with the white dashed lines, reveals that the phase contrast provides different information. The lowest phase values are noticed on a domain whereas on the matrix the changes are between two higher phase levels. These variations are consistent with the assignment of the domain to PS material, which has a high elastic moduli compared to LLDPE that presumably forms the matrix. The phase contrast of the matrix is caused by the probe interactions with lamellar structures and amorphous polyethylene, with the lamellae seen darker.[1] Based on this interpretation we can also assign a number of small inclusions with brighter rims in the PS domains to LLDPE material. This example demonstrates how phase imaging differentiates the components in the multicomponent polymer samples. However, due to the complexity of the phase contrast its assignment to the particular mechanical or adhesive properties is quite difficult.

Block copolymers, another type of polymer materials, are formed of macromolecules that have chemically dissimilar blocks covalently bound to each other. The microphase separation, which minimizes the unfavorable intermolecular interactions, is common for such materials, and it is characterized by periodic patterns with spacing in the 5–100 nm range. AFM is a convenient method for characterization of the structural organization of block copolymers. The tip-force dependent studies of thin film of triblock copolymer poly(styrene)–block–poly (butadiene)–block–poly(styrene) SBS, which was freshly prepared by spin casting on a Si substrate, are instructive for understanding the peculiarities of AFM imaging of block copolymers. The height and phase images of the SBS film underwent reversible changes, when the initial A_{sp}, which was chosen for low-force imaging ($A_{\mathrm{sp}} = 0.9\ A_0$), was reduced twice and then raised back to the initial level, Fig. 7.9a, b. In the height image the force increase converts a relatively smooth surface profiles with small dark depressions to a pattern with dark extended and spherical nanoscale features separated by the brighter spacing. Such a pattern is typical for a microphase separation in SBS with the individual components assembled in the cylindrical blocks with different orientation. The phase contrast variations are even more drastic. At low-force imaging the phase contrast was weak with the features mostly related to the edge effects at the locations related to the depressions. With the force increase the image converts to the microphase separation pattern with the contrast inverse to that of the height image. These changes are consistent with the fact that rubbery poly(butadiene) material, which has lower surface energy than PS, covers the surface in a thin layer. Therefore, at low force the scanning tip profiles this top layer, and at higher force the tip penetrates this layer and starts to interact differently with rubbery poly(butadiene) and glassy PS blocks [19]. Therefore, the PS blocks become elevated in the height image and exhibit a darker contrast in the phase image. The reversible character of

[1] The phase contrast is defined differently in the scanning probe microscopes of different manufacturers. In the NTEGRA microscope made by NT-MDT (Zelenograd, Russia), which was used for most of the measurements presented in this chapter, the phase changes through the resonance from −90° to +90°.

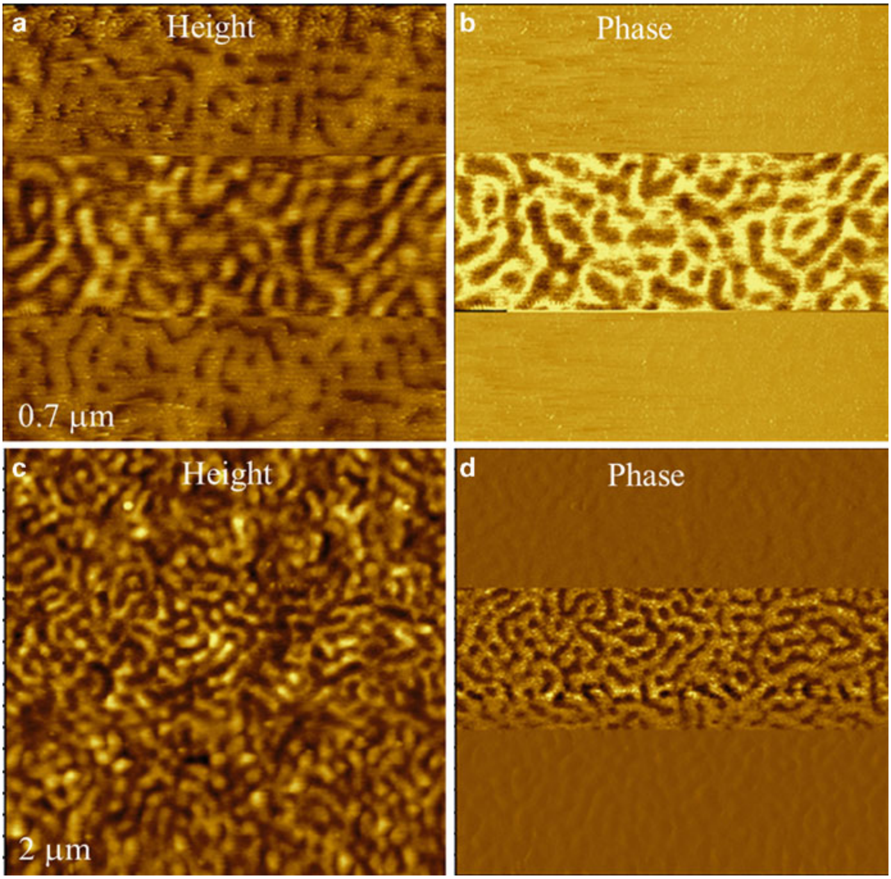


Fig. 7.9 Height and phase images of the films of triblock copolymer SBS in (**a**, **b**) and diblock copolymer PS-PMMA in (**c**, **d**). The imaging was performed in the downward direction and the tip–sample forces were at the elevated level in the middle part of the scans

the tip-force induced changes in the images of the SBS film is explained by a fast recovery of the top rubbery layer.

The variable force imaging applied to the film of diblock copolymer of PS with poly(methyl methacrylate) (PMMA) is reflected in Fig. 7.9c, d. At low tip-force the height images resemble the microphase separation morphology common to this material, whereas the related phase contrast is very weak. The situation changes with a force increase and both images exhibit an identical microphase separation pattern with a reversal of the contrast. The elevated structures in the height images correspond to the darker contrast features in the phase image. From one side, this behavior has some correlations with the changes observed in the SBS film. However, in PS-*b*-PMMA film both components are in the glassy state at RT and their elastic moduli are quite similar. Therefore, the nature of the contrast in this case is still in question.

7.3.4 High-Resolution Imaging

High-resolution imaging is still the most attractive and unique AFM feature, which demands the extensive attention from the instrumental and analysis viewpoints. We will illustrate the related issues on several types of samples. Normal alkanes (chemical formula C_nH_{2n+2}) are linear molecules with a preferential zigzag conformation of the –CH_2– groups. The terminal –CH_3 groups are slightly larger than –CH_2– groups and are more mobile. At ambient conditions the alkanes with $n = 18$ and higher are solid crystals (the melting temperature of $C_{18}H_{38}$ is 28 °C) with the chains oriented practically vertical to the larger faces of the crystals. Such a surface of a $C_{36}H_{74}$ crystal, which is formed of –CH_3 groups, was examined in contact mode, and the AFM images revealed the periodical arrangement of these groups [20]. It has been known for a long time that on the surface of graphite the alkane molecules are assembled in flat-lying lamellar structures, in which the fully extended molecules are oriented along three main graphite directions, Fig. 7.10a. Such molecular order is characterized by a number of periodicities: the 0.13 nm spacing between the neighboring carbon atoms, the 0.25 nm spacing between the –CH_2– groups along the chain in the zigzag conformation, and the 0.5 nm interchain distance inside the lamellae and the lamellae width—the length of the extended C_nH_{2n+2} molecule. The latter varies from 2.3 nm for $C_{18}H_{38}$ to 49.5 nm for $C_{390}H_{782}$ (the longest alkane synthesized). The height and phase images of $C_{36}H_{74}$ alkanes on graphite are presented in Fig. 7.10b, c. The height image shows a layered arrangement with the top layer having 0.5 nm thickness that matches the size of the individual alkane chains. The striped pattern of the layers, which is best seen in the phase image and in the cross-section profile below, has a pitch of 4.5 nm. This dimension is close to the length of $C_{36}H_{74}$ molecules. Therefore, the observed morphology is consistent with the lamellar order of the alkane molecules and the contrast changes reflect the edges of the individual lamella.

The above spatial resolution demonstrated in AM studies of alkane adsorbates on graphite is inferior to that demonstrated in STM imaging of similar samples [21]. The STM images of normal alkanes on graphite clearly demonstrate the fine details of the molecular arrangement such as the lamellar edges, individual chains inside lamellae, and the zigzag conformation of the alkane chains. The visualization of the same features with AFM is not yet achieved in routine measurements. However, the individual chain molecules have already been resolved in the recent studies of dodecanol adsorbates on graphite with the FM mode in liquid [22]. This result is very encouraging, and ongoing instrumental developments aimed at the increase of the signal-to-noise ratio will make such imaging routine.

The ultimate resolution achieved in AFM is often demonstrated by imaging of the atomic-scale and molecular-scale lattices of mica and other layered compounds. Such atomic-scale lattices are easier to get in contact mode as shown by the examples in Fig. 7.11a–c. The images of graphite, MoS_2 and mica exhibit hexagonal patterns with a spacing of ~0.25 nm, 0.32 nm, and 0.52 nm that are close to the crystallographic arrangements of the basal planes of these crystals within the limits

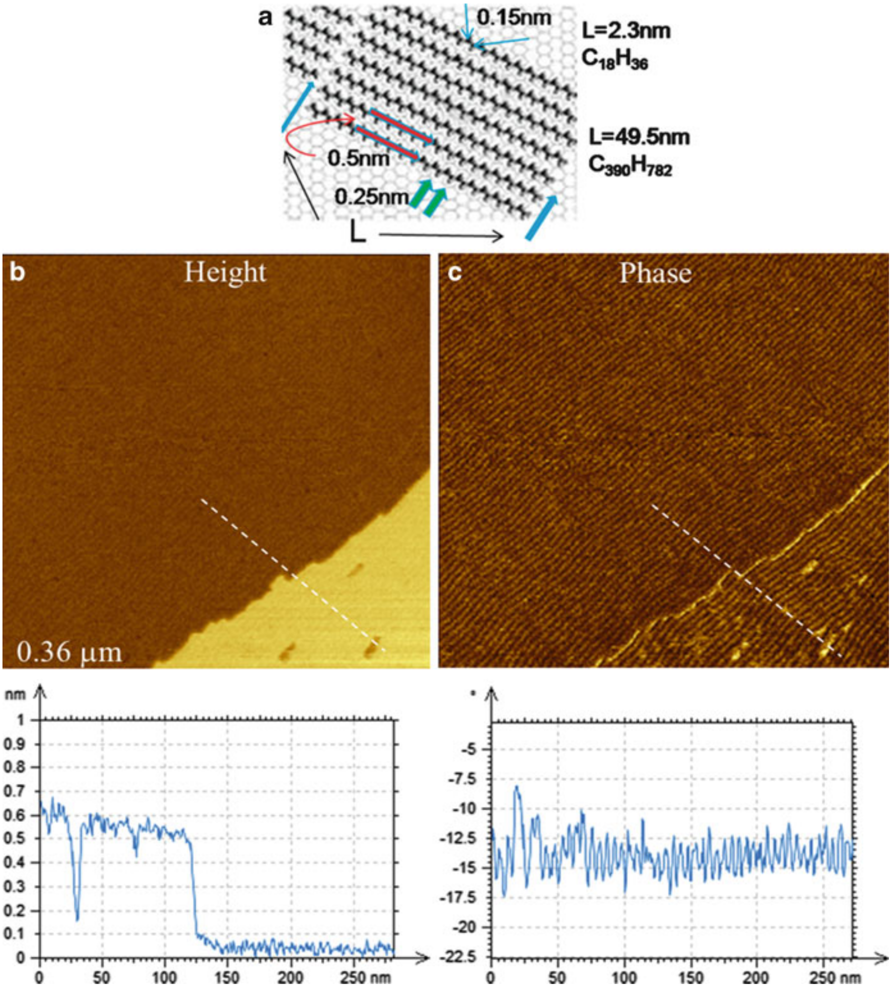


Fig. 7.10 (**a**) Sketch of the lamellar order of normal alkanes on graphite. Height and phase images of normal alkanes $C_{36}H_{74}$ on graphite are in (**b**, **c**). The cross-section profiles taken along the direction marked with the *white dashed lines* are shown underneath the images

of AFM accuracy. The latter is defined by the precision of the piezoelectric scanners, which are typically calibrated with ~5% error. The sketches below the images represent the surface atomic order of these materials, and it is evident that the smaller and tighter spaced C atoms of graphite makes this crystal the most difficult for high-resolution imaging. Indeed, the height corrugations detected in the AFM images of graphite are smaller than those found in the images of MoS_2 and mica. It was noticed that AFM images of a large number of crystalline samples, which were examined with contact mode, exhibit lattices without any single-atom defects. This was considered to be a result of lattice imaging that can be achieved on

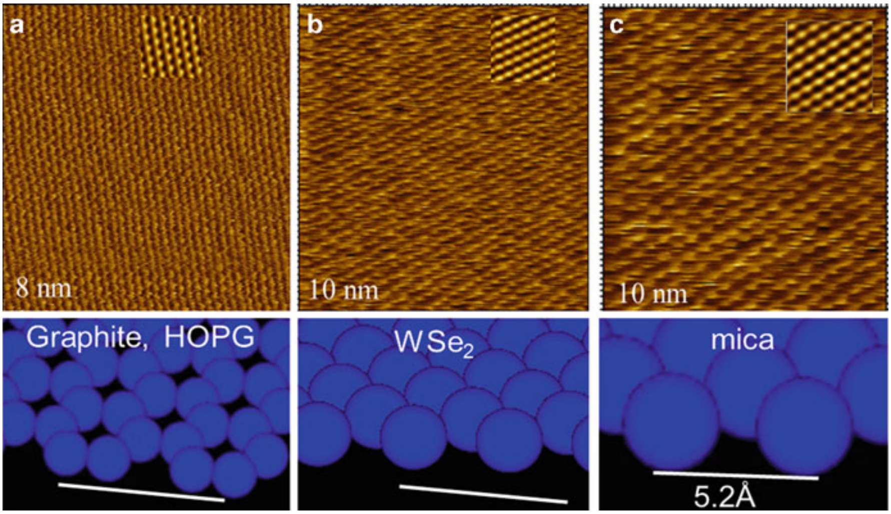


Fig. 7.11 Height images of the surface lattices of graphite, WSe_2, and mica crystals are shown in (**a**), (**b**), and (**c**), respectively. The images were obtained in contact mode. A part of the images at the *top* is presented after FFT-based filtering that emphasizes the periodic structures of the surfaces. The models of the atomic structures of these crystals are shown underneath the images. The *white bar* of 5.2 Å is placed for comparison with the dimensions of the lattices

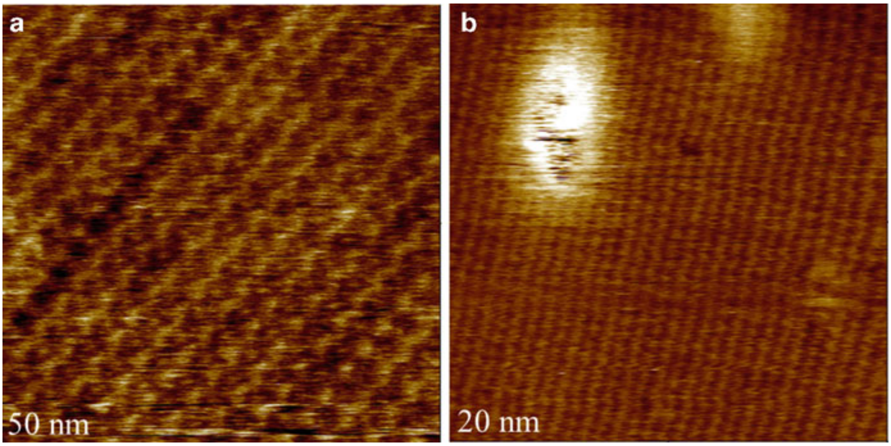


Fig. 7.12 Height images of the molecular structure of a liquid crystalline sample of star-shaped mesogen [21] in (**a**) and polydiacetylene crystal [20] in (**b**). The images were obtained in AM mode

the periodic surfaces with a probe whose tip is larger than the individual atoms. Recently, the observation of the lattices with single defects was achieved in AM and FM mode studies of a number of crystals at ambient conditions and in water [10, 23]. The examples of high-resolution imaging with the AM mode in air are shown in Fig. 7.12a, b. The first image shows a molecular order in the liquid

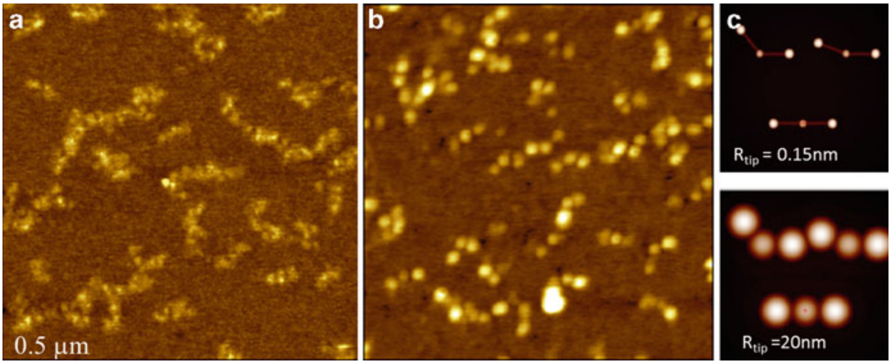


Fig. 7.13 (**a**, **b**) Height images of fibrinogen molecules on mica obtained in AM mode. (**c**) The computer simulated images of fibrinogen molecules in different conformations for a tip with a radius of 0.15 nm and 20 nm

crystalline polymer with multiple structural details at the nanometer scale [24]. A number of the atomic-scale defects are noticed in the height image of the PDA crystal, Fig. 7.12b [23]. These images were obtained in the microscope with an extremely low thermal drift (~0.3 nm/min). Such a low drift allows slow scanning (~1–2 Hz) on the areas below 50 nm, which leads to the improved signal-to-noise performance. The observations of lattices with defects are quite challenging, and at the moment they are more relevant for the demonstration of the high signal-to-noise ratio achieved with a particular AFM instrument rather than examples of true atomic-scale visualization of small objects. This statement is supported by the computer simulation of lattice imaging and visualization of the single-atoms defects, Fig. 7.4a–c. The conclusions drawn from the computer simulation of lattice imaging are quite trivial: although the use of atomically sharp probes and operation with small forces are the main requirements for AFM studies with true atomic resolution, their practical realization necessitates further instrumental efforts.

Visualization of individual macromolecules with AFM is another topic of high-resolution imaging. Successful observations of DNA and polymer molecules on flat substrates [25, 26] give rise to a large area of research. The specific issues of AFM studies of single objects have been partially revealed in Fig. 7.3a–f. In the case of chain molecules, visualization of the molecule shape and conformation, as well as measurements of the contour length, provide unique molecular features that are practically not accessible to other characterization techniques. The height and width of the single objects is more susceptible to limitations related to tip size and tip-force effects. In many cases such effects are inevitable but can be predicted with computer simulations. Important questions related to the conformation of macromolecules of fibrinogen protein on mica in different environments and the structure and morphology of fibrin fibrils were figured out in a recent AFM study [27]. Figure 7.13a, b present the height images of the individual fibrinogen molecules, which are formed of three nodules linked with α-helix coils and have a total length of 47 nm. The fibrinogen molecules and their arrays in Fig. 7.13a resemble the patterns of protein molecules obtained with the sharpest tips in which the subtle details of the nodules are resolved [27]. The image in Fig. 7.13b shows the individual molecules having the tri-nodular shape and the expected contour length. Most of these molecules adopted the bent conformation. Statistical analysis of the images with more than 300 individual fibrinogen molecules on mica showed that straight structures are rare. The molecules are bent at different angles with most probable conformations characterized by bending angles of 106° (27% of molecules) and 157° (59% of the molecules). The results of the computer simulation of AFM images, which were performed for single fibrinogen molecules on mica for the probes with the 0.15 nm and 20 nm tip radii, are shown in Fig. 7.13c. The tip with a 0.15 nm radius represents a single-atom tip, whereas the tip with a 20 nm radii is more practically relevant. The fibrinogen molecules were modeled in the most common bent conformation according to the protein dimensions in the crystal. The simulations were performed in the AM mode with a minimal tip–sample force. As expected, a substantial thickening of the molecular pattern and its transition to the tri-nodular shape occurred when the 20-nm tip was applied. Therefore, it is reasonable to assume that the structure of fibrinogen molecules adsorbed on mica is similar to those found in TEM studies and reconstructed from the X-ray data, and their increased dimensions in the AFM images are the result of tip convolution effects.

7.4 Examination of Mechanical Properties at the Submicron Scale

7.4.1 Force Curves in AFM

Monitoring and adjustments of the tip–sample force interactions are important features for optimization of surface imaging and studies of local mechanical properties of materials [28]. The basic information about these interactions is obtained from force curves, which represent the dependence of the cantilever deflection on the tip–sample distance (*D*v*Z*) measured at a particular surface location. The measurement of force curves in AFM has a similarity with studies using a surface force apparatus [29, 30] and to recording of indentation curves with stylus indenters [31]. Initially, the *D*v*Z* curves were applied for the control and minimization of the normal and lateral forces in attempts to reduce the damage of soft samples. Particularly, the force curves obtained on many samples in ambient conditions manifest a strong capillary force that increases the tip pressure applied to the sample. This meniscus effect, as judged by changes of force curves, can be avoided when a sample and tip are placed under water [28]. The use of force curves for studies of local mechanical properties has substantially broadened AFM applications [32]. The high force sensitivity of this technique inspired a large number of deformation studies, in which individual macromolecules can be stretched by pulling the probe whose apex has been modified to stick to these objects. Such studies are aimed at evaluation of the strength of individual molecules, studies of the intermolecular interactions between the molecular-size objects and exploring sample adhesion. On the other side, by pressing the tip into a sample material we can evaluate its mechanical properties such as the elastic modulus. In addition to unique force sensitivity in such measurements, the small size of the probe was essential in making indentation experiments with nanoscale spatial resolution. In AFM-based nanoindentation the *D*v*Z* curves are collected at rates in the 0.01–10 Hz range, which is well below the resonant frequency of the probes. It is worth noting that the studies are mostly performed on soft materials with an elastic modulus below 10 GPa. This limitation is related to the stiffness of the probes that is below 100 N/m. Stiffer materials are examined with stylus indentors, which are frequently called nanoindentors because of the nanometer scale of the vertical displacements used in these instruments. In other words, the AFM-based nanoindentation and the stylus-based nanoindentation are complementary techniques. The first one is more applicable for the experiments with low forces and higher spatial resolution, whereas the other one covers applications for larger scales and a stiffer range of materials.

In the expansion of local measurements to mapping, force curves can be collected in a mesh of points (up to 128 × 128 points in the operation known as Force Volume) covering a particular surface area. This procedure, however, is time consuming and demands the low thermal drift of the microscope. Such process can be sped up if only a few points of the *D*v*Z* curves are collected for evaluation of stiffness and adhesion maps as it was realized in the pulsed force mode [33]. Further improvements of this mode allowed increasing the force curves harvesting rate to the 1–2 kHz range that brings the mapping of mechanical properties close to the regular imaging speed (~1 Hz).

Several problems of contact mode imaging of soft samples were overcome in the oscillatory AM and FM modes. As we demonstrated earlier, the phase imaging in the AM mode provides compositional mapping of multicomponent polymer materials, which is primarily based on dissimilarities of local mechanical and adhesive properties of their constituents. Although the phase contrast is efficient in differentiating the rubbery, glassy, and inorganic components of polymer blends, block copolymers, and composites, its interpretation in terms of specific mechanical properties is not feasible. To facilitate this problem the recording of *D*v*Z* curves in AM mode was realized in two different approaches. It was shown that these curves can be reconstructed from tip-force responses at different harmonics when the probes with a specific T-shape geometry are applied [34]. A direct detection of the force curves during the AM operation was realized by the use of cantilevers with an interferometric high-bandwidth force sensor [35]. Therefore, the *D*v*Z* curves in the AM mode will be recorded in the tens and hundreds kHz range. The fast recording of force curves with the improved pulsed force and oscillatory modes can be followed by online calculations of the elastic modulus and adhesion if the appropriate deformation models are available. Therefore, the maps of elastic modulus and work of adhesion can be presented together with the height images.

The quantitative analysis of the force curves even in the case of homogeneous materials is rather intricate. The realized recording of force curves at AM mode rates extended to tens and hundreds kHz range provides a possibility to examine the mechanical properties of viscoelastic materials, although this opportunity has not yet been realized.

7.4.2 Towards Quantitative Nanomechanical Measurements

The expectations that AFM instrumentation has reached an automation level capable of providing quantitative results quickly and easily are not implemented so far. This is particularly true in obtaining the mechanical properties at small scales. Such quantitative measurements are possible, but require a thorough approach during the experiment and analysis. We shall consider below the essential elements of nanomechanical studies with AFM and present a number of illustrative examples.

Characterization of the AFM probe is an important step in all nanomechanical measurements. A researcher should undertake several efforts in the characterization of the AFM probes and evaluation of the cantilever optical sensitivity before collecting a reliable set of force curves and imaging the possible indents. Finally, one should choose the appropriate deformation model for extraction of the elastic modulus and work of adhesion. We shall describe these steps below.

The micro-fabrication procedures applied for manufacturing of the AFM probes are not perfect, and variations of the probe geometry are common. Therefore, it is quite unreliable to use the average probe parameters supplied by a manufacturer for quantitative mechanical measurements. Therefore, it is quite desirable to measure the AFM probe stiffness as well as the tip size and apex dimension. Most of the commercial microscopes are equipped with a procedure allowing the measurements of the probe stiffness using the thermal tune method [36]. This method gives most reliable results for soft probes with a stiffness below 5 N/m. Its accuracy is worse (~20% error) when stiffer and short probes are evaluated.

The tip dimensions are among the important parameters in the deformation models applied for the analysis of AFM force curves. The simplest and reasonable assumption is the parabolic shape: $h_{\mathrm{tip}}(r) = r^2/2R$, where R is the tip radius and r is abscissa of the tip profile. At small penetrations it is asymptotically equivalent to the spherical profile: $h_{\mathrm{tip}}(r) = R - \sqrt{R^2 - r^2}$. Formulas for elastic interaction of spheres (Hertz [37]) are based on this asymptotic approach. An additional advantage of the spherical profile is its extension for larger r, although formulas for elastic interaction (Hertz, Sneddon [38], etc.) may not be formally extendable because of the limits of the basic linear theory of elasticity. The main benefit of the parabolic tip profile is its simplicity, but its applicability is best for small penetrations. For large penetrations the complete tip shape should be used.

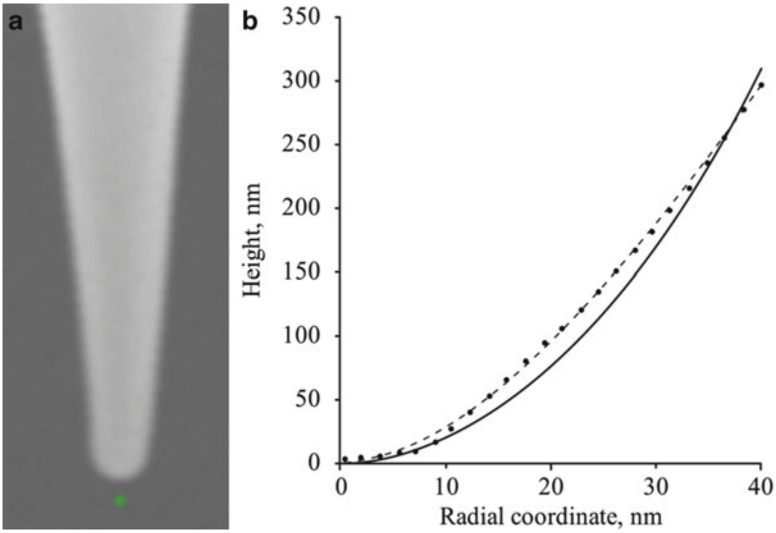


Fig. 7.14 (**a**) A micrograph of the Si tip obtained with TEM microscope. (**b**) The plot of the tip profile. Original points on the tip profile are plotted as *black dots*. The *dashed curve* presents the tip profile approximation with the *n–q–s* function. The *solid curve* reflects the parabolic approximation with optimally fitted radius *R*

The most reliable characterization of the tip shape is achieved with TEM, Fig. 7.14a. For practical use, TEM micrographs of two perpendicular tip projections can be converted to the approximation of the tip-shape function $y_i{=}h_{\mathrm{tip}}(r_i)$. For quantitative analysis we must fit the discrete points ($y_i{=}h_{\mathrm{tip}}(r_i)$) to one of the models of the tip. The choice of the model is a compromise between accuracy and consequent ease of solving the equations to estimate material properties in the framework of one of the solid state deformation theories. The tip profile can be expressed as the linear interpolation (LIT) of given points (r_i, $h_{\mathrm{tip}}(r_i)$) in a piecewise linear form [39, 40]. The latter is well suited to the solution of the Sneddon integrals [38] describing the tip–sample interaction of the arbitrary axi-symmetrical tip with the plane surface. The LIT is the most applicable model for the tip profile although a significant number of the points may slow down the calculations.

Alternatively, the tip profile can be expressed with the *n*-quadratic-signomial (*n*-q-s) function

$$h_{\mathrm{tip}}(r) = c_1 r^n + c_2 r^{2n}, \tag{7.14}$$

where c_1, c_2, and n are fitting parameters; n may not be an integer (since the name “signomial” is in contrast to “polynomial” where n is integer). The advantages of this model are related to a small number (only 3) of fitting parameters with a reasonably good approximation and fast Sneddon calculations by solving linear or quadratic equations [40]. In the simplified form ($c_2 = 0$) the model is used for

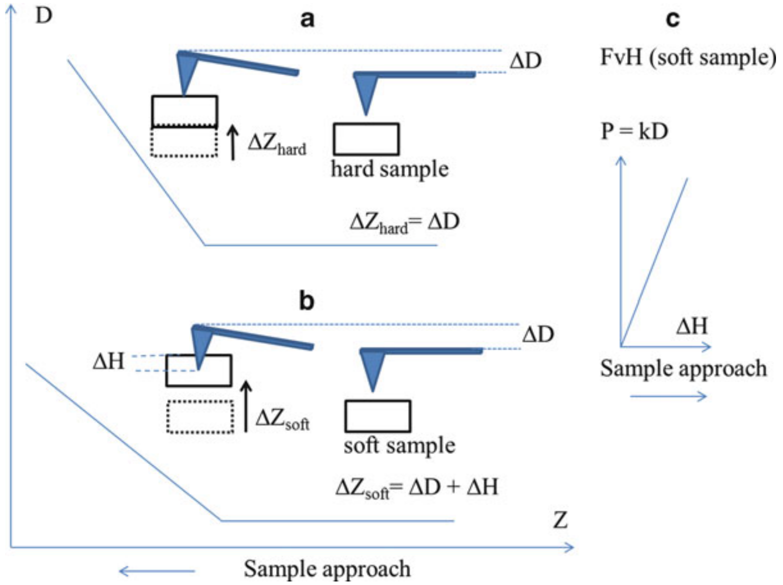


Fig. 7.15 A sketch illustrating the *D*v*Z* dependencies when the AFM probe presses the hard sample (**a**) and soft sample (**b**). The *D*v*Z* curve obtained on the hard sample is used for the estimation of the optical sensitivity ($\Delta D/\Delta Z_{\text{hard}}$). The differences of slopes of the *D*v*Z* curves obtained on hard and soft samples reflect the indentation of the tip into the soft sample—$\Delta H = \Delta Z_{\text{soft}} - \Delta D$. The force versus penetration curve *F*v*H* (**c**) is obtained from the optical sensitivity and ΔH

many analytical developments including those related to the Oliver–Pharr method [41]. The illustration of different fits of the experimental tip shape (Fig. 7.14b) points out that the parabolic approximation is most suited for low penetration experiments; LIT provides the best fit of the experimental points of the tip profile and the *n*-q-s function is a trade-off for the above tip presentations. It is worth noting that the use of the relatively dull probes (Fig. 7.14a) is preferable for nanoindentation experiments. As compared to sharp tips, which are applied for high-resolution imaging, the dull probes exhibit less wear and they are easier for shape characterization.

The optical sensitivity in the optical lever detection scheme defines how the vertical tip motion is recalculated from the photodetector signal that monitors the cantilever bending. This knowledge is needed for getting the probe stiffness from the thermal tune and for the extraction of the tip penetration into soft materials from the force curves, Fig. 7.15a–c. By using the cantilever stiffness and optical sensitivity, the *D*v*Z* curves can be transformed into the *F*v*H* (force-versus-penetration) curves, which are directly related to the mechanical properties of the sample. The optical sensitivity of the probe is directly related to its placement in the holder and to the position of the laser beam on the cantilever surface. Therefore, the use of the same probe for indenting of the sample and calibration purposes is quite important. The optical sensitivity is directly determined from the *D*v*Z* curves obtained on a rigid sample such as diamond or sapphire. In such cases, the tip penetration is negligible and the slope of the force curve defines the optical sensitivity. One should realize that the recording of the force curve on rigid surfaces may lead to tip damage, and such experiments should be made after the measurements of force curves on the sample of interest.

Before performing AFM-based indentation it is quite desirable to examine the surface morphology of the sample and to choose the particular locations. For homogeneous samples the indents performed in different locations are necessary for getting numerous *D*v*Z* curves for averaging. In case of heterogeneous materials the preliminary imaging will allow indenting the individual components or specific locations for further comparison of their properties. It is also important to perform the measurements at different force levels by triggering the maximal probe deflection. Such measurements can be invaluable for separation of the elastic and plastic contributions to the force curves, which will simplify their quantitative analysis. In some instances it is useful to record the force curves at different rates and figure out the possible time-dependent (viscoelastic) effects. After the force curves are collected it is worth reexamining the same locations to visualize the possible indents left and to measure their shapes and dimensions. The additional information, such as visualization of the possible pile-ups around the indents, will be invaluable in the analysis of the local mechanical properties. Imaging of the sample locations either before or after recording the force curve is better to be done in AM mode and not in contact mode. This will help avoid a possible damage of the sample or modification of its surface that can deteriorate the reproducibility and reliability of the force curve experiments.

For the illustration of AFM-based nanoindentation studies we chose the results obtained on a flat surface of a LDPE block, Fig. 7.16a–c. They include a characteristic *D*v*Z* curve, the height image of the nine indents' location, and the cross-section profiles taken along the indents. In the nanoindentation experiment the probe is moved to the sample and, after the tip touches it, the *D*v*Z* curve is gradually rising until the probe reaches the trigger level, which was 10 nN for this experiment. After the loading stops, the probe is moved back to the rest position, and the unloading curve is quite different from the loading one. This dissimilarity is the consequence of energy dissipation caused by an inelastic nature of the sample deformation made by the probe. Furthermore, in this indentation process the tip left the sample at more advanced *Z*-position ~250 nm compared to the initial point of the contact. This difference defines the depression left by the indenting tip. The depth of the indents in the image is much smaller and this points to a fast recovery of a substantial part of the depression. Therefore, the tip-induced indentation of LDPE material is accompanied by viscous effects and remaining plastic deformation. Earlier, we have recorded the time-dependent recovery of the indents made in the layers of ultra-low density polyethylene [42]. Such behavior is also common for mechanical testing of polymers at the macroscopic scale.

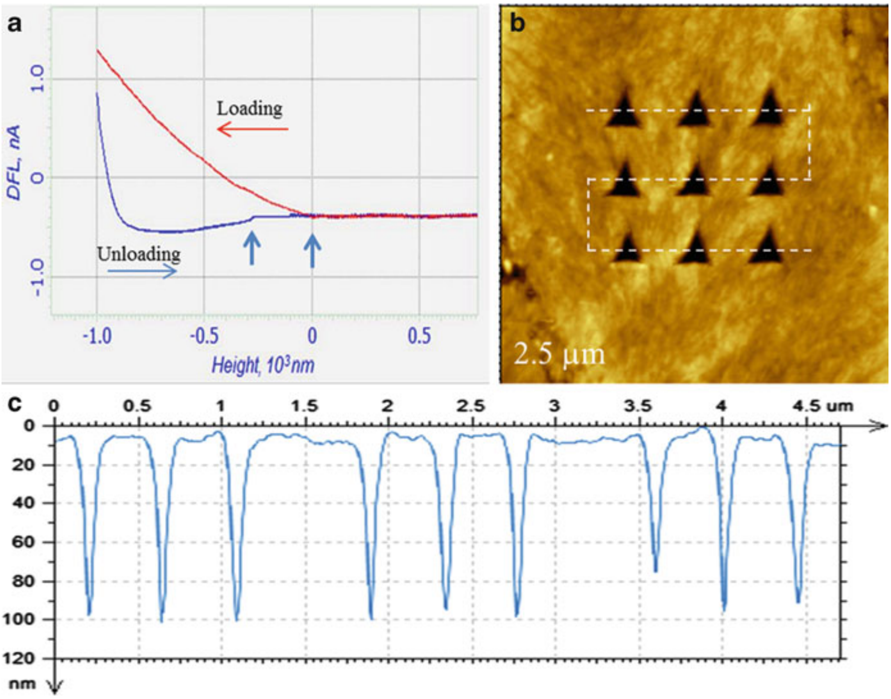


Fig. 7.16 The *D*v*Z* curve (**a**), the height image of the indents made in LDPE sample (**b**), and the cross-section profile (**c**) taken along the direction pointed out with the *white dashed line*

This example shows that AFM-based indentation of polymers is a rather complicated phenomena and the extraction of valuable mechanical properties from such experiments is a challenging task. So far, the development of quantitative nanomechanical measurements in AFM-based nanoindentation has been performed on homogeneous polymers. This simplifies the collection of reliable and reproducible *D*v*Z* curves that need to be analyzed. The measurements of atactic PS can be considered as an example of such studies, Fig. 7.17a–f. The force curves, which were obtained on this polymer at different maximal loads (50 nN and 250 nN), are presented together with the images taken at the indented locations. At the small load (Fig. 7.17a), the loading and unloading traces follow each other, which indicates the elastic nature of the deformation. The presence of attractive force regions close to the point of the initial contact is a consequence of the tip–sample adhesion. As seen from the height image in Fig. 7.17b, the surface area, where the indents were made, did not exhibit any indentation marks. The situation is different when the tip-force is increased. The related force curves (Fig. 7.17c) have different loading and unloading parts, and the indents with small pile-ups around become visible in the height image (Fig. 7.17f). The remaining indents were around 3 nm in depth. The *F*v*H* curve in Fig. 7.17d is obtained from the *D*v*Z* curves in Fig. 7.17c, and it is the subject of the quantitative analysis.

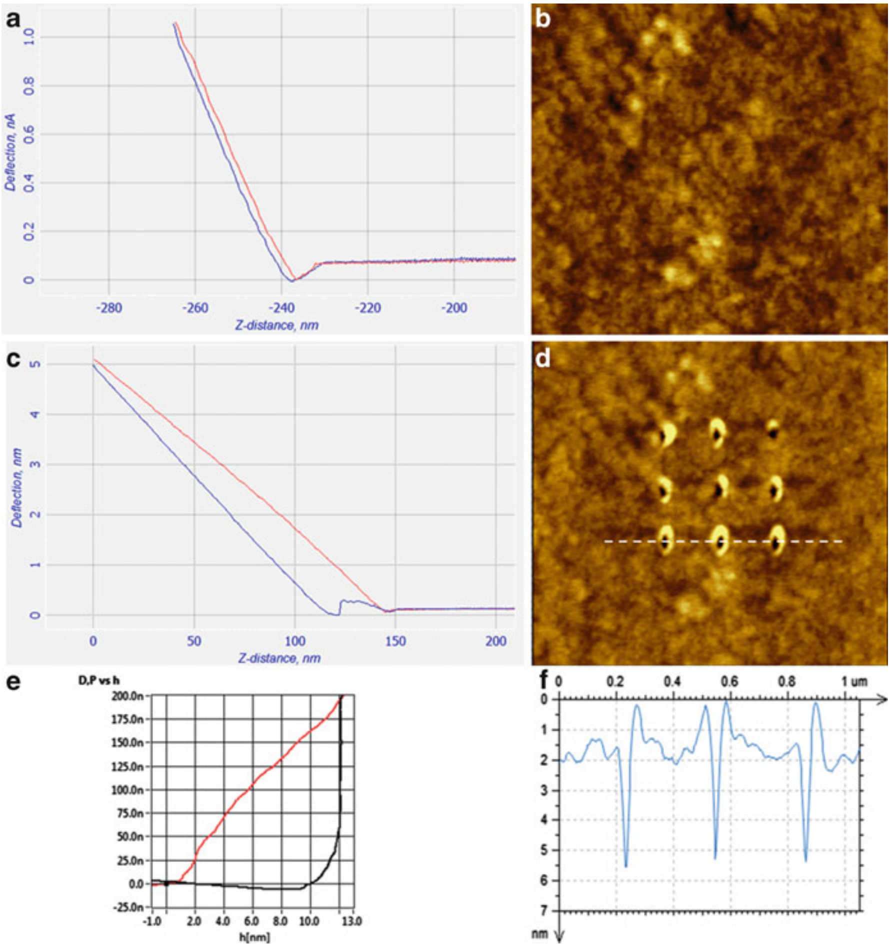


Fig. 7.17 The *D*v*Z* curves and the images of the area where the AFM-based indentations were made on a PS sample with the forces of 50 nN (**a**, **b**) and 250 nN (**c**, **d**). The graph in *E* presents the *F*v*H* curve recalculated from the *D*v*Z* curve in (**c**). The cross-section profile along the three indents in the direction shown with a *white dashed line* in (**d**) is presented in (**f**)

7.4.3 Theoretical Models and Analysis of AFM Force Curves

The final and most invaluable step in AFM studies of local mechanical properties is the analysis of *F*v*H* curves in terms of a properly chosen model. Although far from complete, it is recognized that the list of the most important mechanical properties of materials includes *VEPA*: here *V* is the viscosity, *E* the elasticity, *P* the plasticity, and *A* the adhesion. Although theories covering these individual properties have been developed, there are no justified synergetic models of *VEPA*. Besides, these properties may not be uncorrelated, which makes it difficult to separate them.

We will consider a possible analysis of the AFM-based nanoindentation results using the above data obtained on PS.

As regarding the elastic models, which can be applied for the force curves shown in Fig. 7.17a, we describe the elastic model of Hertz [37] and its generalization for arbitrary tip profile derived by Sneddon [38]. The elastic Hertz mode, which describes the parabolic tip interacting with a plane surface, has the following relationships between the applied force P, the probe parameters (tip radius R), the mechanical characteristics of the sample ($K = 4/3E_\mathrm{r}$, where E_r is the reduced elastic modulus; w the work of adhesion), and the deformation characteristics (a the contact radius and h the tip penetration)

$$\frac{a^3K}{R} = F, \quad h = \frac{a^2}{R} \tag{7.15}$$

The well-known 3/2 power dependence is derived from Eq. (7.15):

$$F = K\sqrt{R}h^{3/2}. \tag{7.16}$$

The Hertz model is applied in the case of a parabolic tip whereas the Sneddon approach is more general by considering an arbitrary axi-symmetrical tip interacting with a plane sample. The related formulas contain the Sneddon integrals:

$$h_\mathrm{S}(a) = \int_0^1 \frac{f'(x)\mathrm{d}x}{\sqrt{1-x^2}}, \quad P_\mathrm{S}(a) = 2aE_\mathrm{r}\int_0^1 \frac{x^2f'(x)\mathrm{d}x}{\sqrt{1-x^2}} \tag{7.17}$$

where function $f(x)$, $0 \leq x \leq 0$, relates to the tip profile by the formula

$$h_\mathrm{tip}(r) = f(r/a), \quad 0 \leq r \leq a. \tag{7.18}$$

Subscript “S” is used in h_S and F_S to emphasize that these are calculated by Sneddon integrals. These integrals can be presented in closed form for the tip-shape models described earlier. In particular, for the n-q-s function, a (h) can be derived from the first expression in Eq. (7.17) as a formula for the solution of the quadratic or linear equation. Substitution of the expression for a (h) into the second formula of Eq. (7.17) provides an analytical expression of P as function of h. For the LIT model, the derivative of the function f is piece-wise constant and the integrals in Eq. (7.17) become sums of simple analytical expressions. The Hertz and elastic Sneddon models were implemented in a LabVIEW-based interactive software and applied for the calculation of elastic moduli of PDMS and other homogeneous polymer samples [39, 40].

In addition to elastic models, we will shortly describe the elasto-adhesive and elasto-plastic models. There are two elasto-adhesive extensions to the pure elastic Hertz approach: DMT (Derjaguin–Muller–Toporov [43]) and JKR (Johnson–Kendall–Roberts [44]) models. The case of a parabolic tip interacting with a plane sample is treated by the DMT model that differs from the Hertz model by the subtraction of the adhesive force equal to $2\pi wR$. The corresponding equations are shown below:

$$\frac{a^3 K}{R} - 2\pi wR = F, \quad h = \frac{a^2}{R} \tag{7.19}$$

or

$$F = K\sqrt{R}h^{3/2} - 2\pi wR. \tag{7.20}$$

In this case the constant attractive adhesive force is acting in addition to repulsive elastic force.

In the JKR model both F and h have decreased, as seen from the relationship below:

$$\frac{a^3 K}{R} - \sqrt{6\pi a^3 Kw} = F, \quad h = \frac{a^2}{R} - \sqrt{\frac{8\pi aw}{3K}}. \tag{7.21}$$

In contrast to the DMT model, the subtraction terms depend explicitly on the contact radius a that makes the computations more complex. A sketch showing the differences of the contact area and depression in the Hertz and JKR models is presented in Fig. 7.18a. The adhesion increases the contact area and the tip penetration. The FvH curves for the Hertz, JKR, and DMT model graphs are shown in Fig. 7.18b. These models can be generalized for the arbitrary axi-symmetrical tip profile by using the more general Sneddon integrals (7.17).

Analysis of the deformation curves obtained with stylus nanoindentors is often based on the Oliver–Pharr (OP) method [41]. The pure elastic Sneddon model with a plastic correction is used in this method to treat the dissipative deformation. In this case, the modulus of the elasto-plastic materials is determined from the unloading curves and heuristic plastic correction, which is based on the value h_f called "the residual hardness impression." The OP method has been verified on hard materials such as metals and inorganic glasses that are free from viscoelastic effects. The applicability of this approach to polymer materials was not thoroughly proved.

The analysis of the FvH curves obtained for a PS sample (Fig. 7.17a–f) showed the following results. The force curves of the low-force experiments, which described the elastic deformation, were treated in terms of the JKR model and the estimated elastic modulus was ~2.5 GPa. This result is in line with the earlier data [39, 40] and macroscopic elastic modulus of this polymer. The initial part of the loading part of the FvH curve in Fig. 7.17e was used for the estimation of the elastic modulus with the Hertz model. We also applied the OP method and estimated the elastic modulus employing the loading and unloading traces. In all case, the modulus was in the 2.1–2.5 GPa range.

The given example and the earlier elastic modulus and work of adhesion data, which were obtained on homogeneous polymers [39, 40], are well correlated with

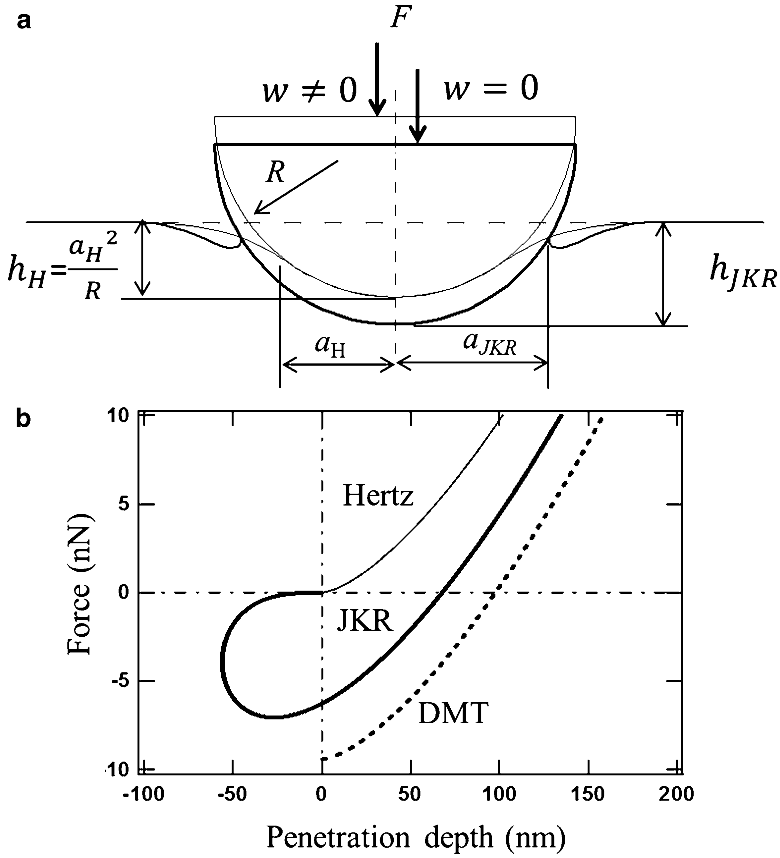


Fig. 7.18 (**a**) Sketch describing the differences in penetrations and contact areas in the Hertz and JKR models, which takes into account the indentor–sample adhesion. (**b**) Graphs representing the force versus penetration curves in three different models: Hertz, JKR, and DMT

the mechanical characteristics obtained on macroscopic samples. This is an important, but only a first step, in the nanomechanical AFM applications, because the main interest is in studies of heterogeneous materials and interfaces in multicomponent systems. The initial efforts in this field [45] revealed that this is a nontrivial task that demands further efforts.

7.5 Measurements of Local Electric and Dielectric Properties

There are two types of local electric methods associated with AFM. Electrostatic force detection is employed in electric force microscopy, Kelvin force microscopy (KFM), and measurements of capacitance gradients (d*C*/d*Z* and d*C*/d*V*). In these techniques the tip–sample forces are used simultaneously for tracking surface topography and for electric measurements. In the other methods, the tip–sample force interactions are applied for profiling whereas the electric property (e.g., current, capacitance) is measured with an additional sensor. Here we will focus primarily on measurements of the surface potential and local dielectric properties through detection of the electrostatic force and its gradient. Such applications are expanding with the development of multifrequency techniques.

7.5.1 Electrostatic Forces in AFM

The sensitivity of the conducting probe to electrostatic force was demonstrated in the first AFM applications [46]. In this work, the 1st flexural mode (ω_{mech}) of the probe resonant oscillation was chosen for tracking the sample topography during noncontact operation. Simultaneously, the electrostatic tip–sample force was stimulated by an AC voltage applied to the probe at a much lower frequency (ω_{elec}). The related changes of amplitude at ω_{elec} reflect the variations of the electrostatic force caused by local surface charges, dipoles, or regions with different work functions and doping type or level. Therefore, monitoring and recording of probe responses at two different frequencies allows simultaneous and independent measurements of local electric and mechanical interactions with the latter applied for surface profiling. This principle is applied to single-pass studies of surface potential in KFM and capacitance gradient dC/dZ (Z is the vertical distance between the probe and the sample). The detection principle of surface potential and dC/dZ is based on equations describing the tip–sample electrostatic interactions using a capacitor-like model [46]. The quadratic dependence of the force on the difference of surface potentials of the probe and sample gives rise to several force components when external DC and AC voltages—U_{DC} and U_{AC} (the latter at frequency ω_{elec}) are applied to the probe to promote electrostatic forces $F_{\omega_{\mathrm{elec}}}$ and $F_{2\omega_{\mathrm{elec}}}$.

$$F_{\omega_{\mathrm{elec}}}(Z) = -\frac{\partial C}{\partial Z}[(\varphi - U_{\mathrm{DC}})U_{\mathrm{AC}}\sin(\omega_{\mathrm{elec}}t)] \tag{7.22}$$

$$F_{2\omega_{\mathrm{elec}}}(Z) = -\frac{\partial C}{\partial Z}U^{2}{}_{\mathrm{AC}}\cos(2\omega_{\mathrm{elec}}t). \tag{7.23}$$

In Eq. (7.22), φ is the difference between the surface potential of the probe and the sample location beneath the probe. The surface potential difference is determined by finding a specific U_{DC} that nullifies the force at ω_{elec}. This is the task of the KFM servo. The capacitance gradient dC/dZ is directly proportional to the electrostatic force at the 2nd harmonic of ω_{elec}, and it is related to the dielectric permittivity (ε) of the material underneath the probe. For many materials the dielectric permittivity can be a complex value and measurements of the real and imaginary components of dC/dZ are essential. Here, we do not consider the dependence of the capacitor on applied voltage, but in a more general case the dC/dV gradient is related to the electrostatic force response at $3\omega_{\mathrm{elec}}$ [47]. Therefore, several lock-in amplifiers, which are tuned to ω_{mech}, ω_{elec}, $2\omega_{\mathrm{elec}}$, and $3\omega_{\mathrm{elec}}$, can enable simultaneous measurements of topographic and various electric and dielectric properties of samples.

Such a multifrequency AFM approach has definite advantages when compared to the well-known two-pass method, in which the measurements of the surface topography and local electric properties are performed in separated passes with the conducting probe being retracted from the surface at a lift distance of 10–20 nm [48]. Despite its simplicity, the two-pass method has a number of limitations. They are related to (a) an undesirable "contamination" of the topography images by electrostatic forces acting between the probe and sample, (b) a problem of finding an appropriate lift height to avoid the mechanical interactions during the 2nd pass over corrugated surfaces, (c) a loss of spatial resolution and sensitivity caused by the distant position of the probe in the 2nd pass. The latter limitation does not exist in the single-pass technique because of a closer proximity of the probe to the sample. The comparative advantages of the single-pass operation have been already demonstrated [49, 50].

7.5.2 Single-Pass KFM and Dielectric Response Measurements

Recent developments of AFM electronics, which enable multifrequency measurements and various oscillatory modes, enhance the researcher's capability in finding the optimal experimental routine for advanced studies of surface properties. The combinations of amplitude modulation (AM) and frequency modulation (FM) were explored for surface tracking and KFM mapping in UHV conditions [51, 52]. The most sensitive and accurate surface potential data was obtained using the electrostatic force gradient detection with FM. These results are consistent with the earlier theoretical estimates [53] that pointed out that in addition to the tip apex, both the cantilever and the tip body contribute substantially to the overall electrostatic force exercised by the AFM probe. These contributions are eliminated when the force gradient is measured. This is true when regular tips of ~10 μm in height are applied. The surface potential data obtained with force and force gradient detections are essentially identical when using conducting probes with extra-long tips (~100 μm in height) [54].

In KFM studies under ambient conditions we have utilized both force and force gradient for measurements of surface potential and dielectric response. While the probe is driven to mechanical oscillation at ω_{mech} and its electric bias with respect to the sample is changing at ω_{elec} ($\omega_{\mathrm{elec}} \ll \omega_{\mathrm{mech}}$), the related frequency components of the photodetector signal can be monitored in parallel with separate lock-in amplifiers (LIA). In this case the sample topography will be tracked based on maintaining the set-point amplitude (A_{sp}) with the 1st LIA, which is tuned to ω_{mech}.

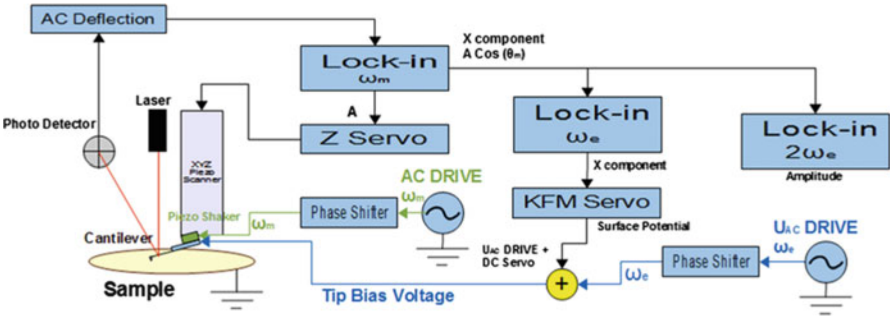


Fig. 7.19 Sketch representing the instrumental set-up in the KFM-PM mode

Simultaneously, the 2nd LIA, which is tuned to ω_{elec}, records the amplitude that is proportional to $F_{\omega_{\mathrm{elec}}}(Z)$. Furthermore, the servo, which is incorporated in the instrument loop consisting of a probe, photodetector, and 2nd LIA, can adjust U_{DC} [see Eq. (7.22)] to nullify the force and thus to determine the local surface potential φ. This operation is often known as KFM-AM, where AM indicates a detection of the electrostatic force $F_{\omega_{\mathrm{elec}}}$. In the parallel connection of two LIA, the tuning of the 2nd LIA to $2\omega_{\mathrm{elec}}$ will enable recording of the dC/dZ signal. The amplitude and phase (or real and imaginary components) of this signal are essential for samples with complex dielectric permittivity. When a 3rd LIA is also added in parallel then 2nd and 3rd amplifiers can detect the responses at ω_{elec} and $2\omega_{\mathrm{elec}}$ thus enabling the simultaneous recording of sample topography, surface potential, and capacitance gradient.

In an alternative way, as shown in Fig. 7.19, the surface potential and $\frac{\partial C}{\partial Z}$ can be measured in the LIA configuration employing 2nd and 3rd LIA, which are connected in series with the 1st amplifier. In this case the electrostatic force is stimulated by U_{AC}, which is applied to the probe at low frequency that is within the bandwidth of the mechanical probe oscillation at ω_{mech}. A combination of mechanical and electrostatic tip–sample interactions will provide the additive contributions to the phase of the photodetector signal: $\theta(t) = \theta_{\mathrm{mech}} + \Delta\theta_{\mathrm{elec}}(t)$. For the best electrostatic performance, mechanical forces should be minimized by setting A_{sp} close to A_0 (the probe amplitude prior to the engagement on the sample). In this case $\theta_{\mathrm{mech}} \approx \pi/2$ and $\cos\theta(t) \approx -\Delta\theta_{\mathrm{elec}}(t)$. This means that the frequency components of the phase signal can be applied for measurements of the surface potential and the capacitance gradient.

It is worth noting that the suggested use of phase modulation of the electrostatic force interactions is quite similar to FM because both methods, under simplifying assumptions, provide the force gradient data. The formalism applied to FM shows that

$$\frac{\Delta f}{f_0} = -\frac{1}{\pi k A}\int_{-1}^{1} F_z[d_{\min} + A(1+u)]\frac{u\mathrm{d}u}{\sqrt{1-u^2}}. \tag{7.24}$$

Using the approximation of small amplitudes, the expression becomes simpler with the following relation between the frequency shift and the force gradient [55]:

$$\frac{\Delta f}{f_0} = \frac{1}{2k}\left[\frac{\partial F_z(z)}{\partial z}\right]_{z=d_{\min}}. \tag{7.25}$$

The relationship between the cosine of the phase and the tip–sample interaction forces, which was shown above [Eq. (7.12)], is valid for the electrostatic forces. Using the approximation of small amplitudes, the cosine of the phase is also proportional to the force gradient:

$$\cos\theta = -\frac{2}{N}\int_0^{\pi} F_z(Z_\mathrm{c} + A\cos y)\cos y\,\mathrm{d}y \approx \frac{\pi A}{N}\frac{\partial F(z)}{\partial z}\bigg|_{z=Z_\mathrm{c}}. \tag{7.26}$$

The use of the cosine of the phase for KFM feedback is more precisely connoted KFM-PM, where PM is the phase modulation by the electrostatic force gradient. In summary, we are using KFM-AM and KFM-PM approaches, which are based on the detection of electrostatic force and its gradient, for measurements of the surface potential. A similar methodology is applied to measurements of dC/dZ, which can be performed in combination with KFM or independently. For our practical applications, the most salient point is that the measurements at different frequencies are performed using the intermittent contact mode with a relatively small level of mechanical tip–sample interactions. Such experiments conducted on various samples demonstrated that any cross-talk between the topography and local electric measurements is essentially absent [56].

7.5.3 Practical Examples of KFM and Dielectric Studies

The use of different signal detection schemes and the development of multifrequency techniques provide the AFM researcher with a variety of methods for the comparison and examination of the same properties. This is the situation with surface potential and dC/dZ measurements. Therefore, studies of standard samples are invaluable for the verification of different techniques and their applicability. For our KFM experiments we selected samples of self-assemblies of semifluorinated alkanes on different substrates (Si wafer, mica, and graphite), semiconductor SRAM structures, and a bimetallic alloy of Bi/Sn.

The AFM images, which illustrate the KFM-PM and dC/dZ measurements of self-assemblies of semifluorinated alkanes $CF_3(CF_2)_{14}(CH_2)_{20}CH_3$ ($F_{14}H_{20}$) on a Si substrate, are presented in Fig. 7.20a–d. The domains of $F_{14}H_{20}$ self-assemblies with spiral features that are less than 4 nm in height are seen in the height and phase images, Fig. 7.20a, b. The fact that the phase contrast visualizes only the edges of domains, which are much softer than the substrate, indicates that the measurement

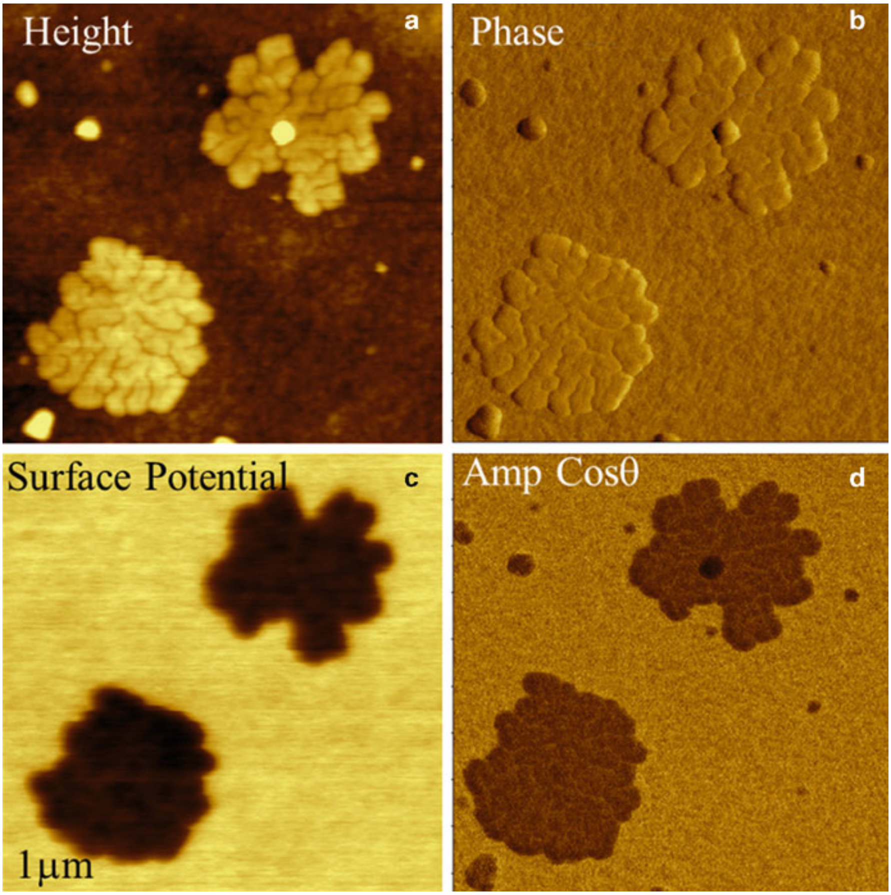


Fig. 7.20 Height (**a**), phase (**b**), surface potential (**c**), and dielectric response (amplitude of cos θ at $2\omega_{elec}$) (**d**) images of $F_{14}H_{20}$ self-assemblies on a Si substrate obtained in the single-pass KFM-FM and dielectric studies

was conducted at low force. As mentioned earlier, this is a main requirement of the KFM-FM measurements. The structure of $F_{14}H_{20}$ self-assemblies on Si reflects the dissimilar molecular nature, conformation, and volume of the fluorinated and hydrogenated parts, which are covalently linked into one chain-like molecule. It is expected that the more bulky fluorinated segments are organized at the exterior of the spirals facing air [57]. This arrangement leads to a preferential vertical orientation of the fluoroalkanes molecules, which have a strong dipole at the central junction –CF_2–CH_2– oriented along the chain. Therefore, a strong negative surface potential of the $F_{14}H_{20}$ domains is expected, which was proven with the macroscopic Kevin probe studies of Langmuir–Blodgett layers of the semifluorinated alkanes [58] and in earlier KFM measurements [49, 59]. The negative surface potential is distinctively seen in the surface potential image shown in Fig. 7.20c.

The negative potential contrast is noticed only at the domains whereas small particles are not visualized in the surface potential image. The lack of small particles in the surface potential data is a strong indication that there is no noticeable cross-talk between the mechanical and electrostatic forces in the single-pass operation. The same particles and $F_{14}H_{20}$ domains are seen in the dielectric response image (Fig. 7.20d); this is expected because any material between two electrodes will change the capacitance gradient.

High sensitivity and spatial resolution can be achieved in single-pass KFM-PM studies as demonstrated in the images of $F_{14}H_{20}$ self-assemblies on graphite, Fig. 7.21a–f. In contrast to other substrates (Si, mica), for the $F_{14}H_{20}$ adsorbates on graphite the first layers are formed of molecules, which are oriented parallel to the substrate and form lamellar structures of 6–8 nm in width. The molecular dipoles in these layers are also preferentially oriented parallel to the surface. Therefore, the surface potential of these locations will be less strong compared to the $F_{14}H_{20}$ self-assemblies on Si. The large-scale height image shows a number of flat lamellar sheets and numerous droplets dispersed between them, Fig. 7.21a. The surface potential contrast of the droplets is moderately negative (around −200 mV), whereas surface potentials of the lamellar sheets and, particularly, of the bare substrate locations are more positive, Fig. 7.21b. This is consistent with the expected orientation of the molecular dipoles parallel to the substrate plane. The high-resolution height and surface potential images in Fig. 7.21c, d revealed lamellar patterns with height corrugations in the 300 pm range and potential changes in the 10–20 mV range, which are pointed out by the cross-section profiles in Fig. 7.21e, f. The spatial resolution of these images, which show the 6-nm spacing, is around a few nm as judged by the width of the dark strips. Additional experimental and theoretical efforts are needed for a complete analysis of this arrangement [59].

As mentioned earlier, AFM-based electric modes can be used to analyze various materials, and such measurements do not suffer the stiffness-related limitations of local nanomechanical measurement methods. Semiconductor SRAM structures were examined with KFM-AM at two locations and the representative images are shown in Fig. 7.22a–d. The surface potential of semiconductor structures depends on type and doping density. Therefore, one should not expect a direct correlation between topography and surface potential images. The surface potential patterns of these semiconductor structures are quite different from the sample topography seen in the height images. Remarkably, the surface potential images of the large-scale SRAM and other semiconductor structures have better stability and resolution when they are imaged using KFM-AM mode. Subsequently, the optimization of KFM measurements on different sample types must include both the proper selection of the imaging mode and the probe type. The use of larger tip radius conducting probes is favored for a higher signal-to-noise ratio when measuring local electric properties, and these probes are also preferred because of their higher wear resistance when very high spatial resolution is not required.

Another example of a rigid sample, which can be successfully examined with KFM is the soldering material, bismuth–tin: BiSn. A specimen of this incomplete

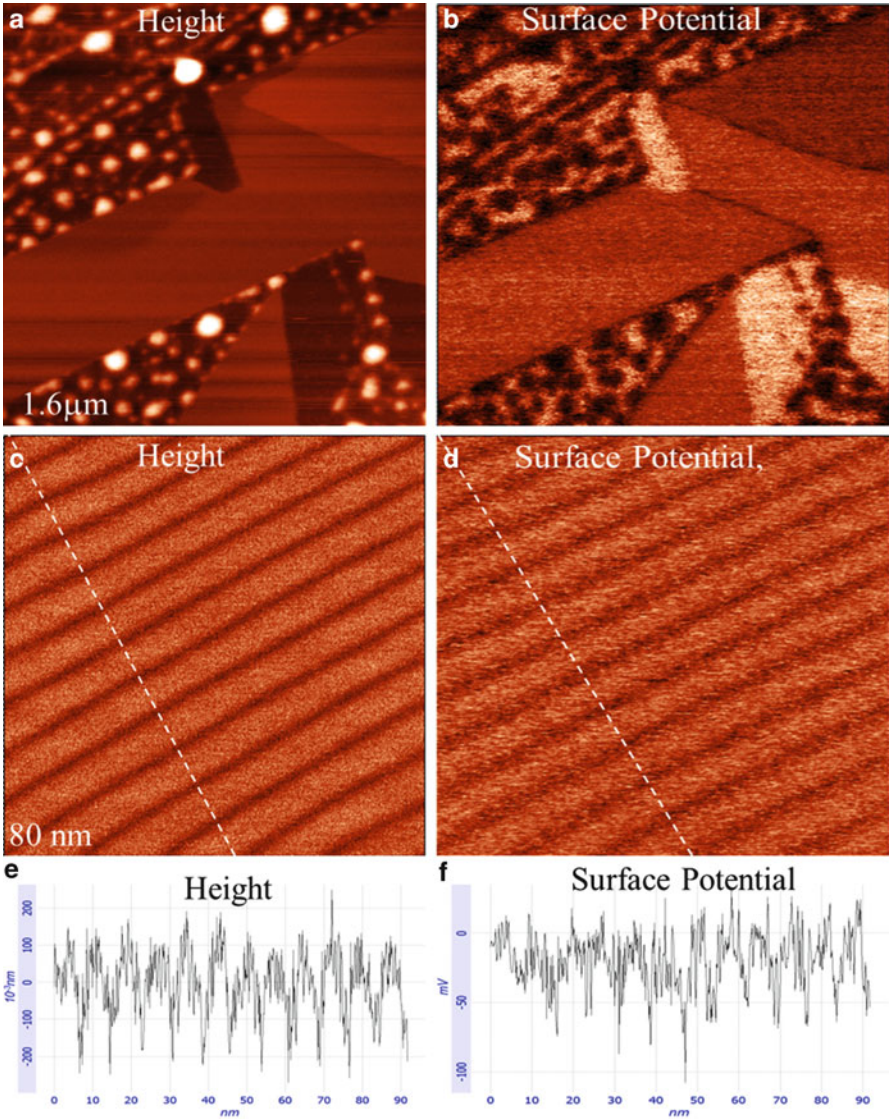


Fig. 7.21 Height (**a**, **c**) and surface potential images (**b**, **d**) of $F_{14}H_{20}$ self-assemblies on graphite obtained in the single-pass KFM-PM mode at two different scales. (**e**, **f**) The height and surface potential profiles are taken along the directions indicated in the images with the *dotted lines*

metal alloy can be prepared for AFM studies as a flat sheet by melting the material between two flat substrates. The height, phase, and surface potential images of these samples reveal a relatively smooth surface morphology with the domain structures separated by 10–20 nm steps, Fig. 7.23a–c. The phase image emphasizes the edges of the domains, whereas the surface potential image exhibits a completely different

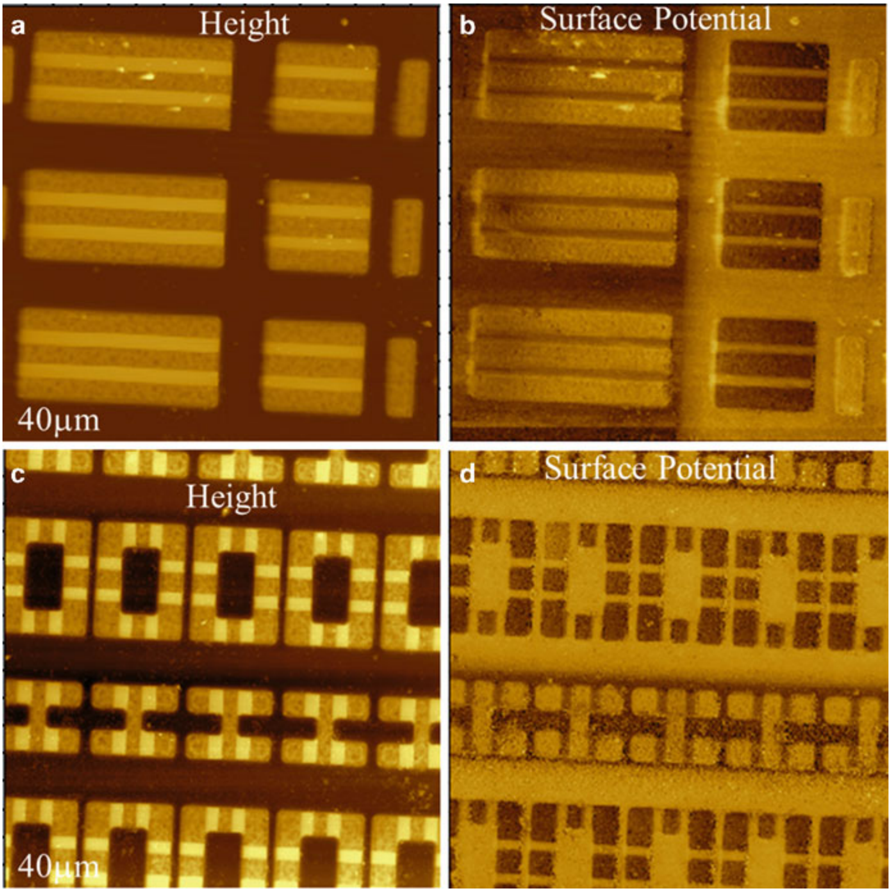


Fig. 7.22 Height (**a**, **c**) and surface potential images (**b**, **d**) obtained at two locations of SRAM structure in the KFM-AM mode

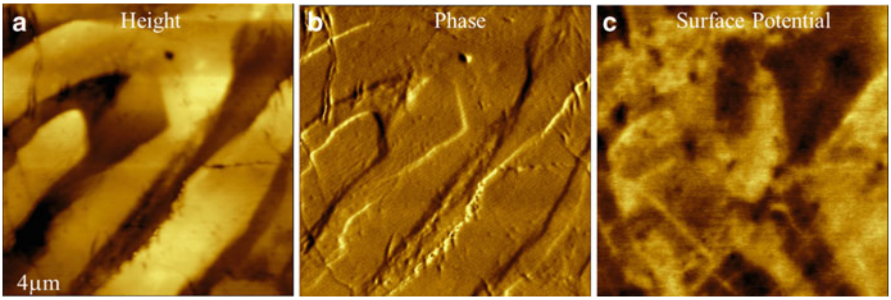


Fig. 7.23 Height (**a**), phase (**b**), and surface potential (**c**) images of Bi/Sn alloy obtained in the KFM-PM mode

pattern that is uncorrelated with the surface topography. The surface potential variations are in the range of 200 mV, which is consistent with the difference in the work functions of Bi and Sn. The binary contrast in the surface potential images of Bi/Sn alloy might deteriorate with oxidation, which is particularly strong for Sn. Surface oxidation can cause the compositional map to become less pronounced [59]. It is worth noting that KFM is actually an exception to common AFM techniques in that it directly provides quantitative values for a particular sample property. In the case of metals, the surface potential is related to the local sample work function, and for molecular systems with dipoles, surface potential correlates with the strength and orientation of molecular dipoles. Kelvin force measurements are also applied for studies of free charges and their behavior caused by various dynamic processes. However, one should not overestimate the capabilities of KFM measurements, particularly when they are performed in ambient conditions. A possible contamination of the sample or the probe might substantially change the validity of absolute values of the surface potential obtained in such studies. Therefore, surface potential differences, which are measured at various locations within a particular scan, are more reliable than the absolute surface potential values.

7.5.4 On the Way to Quantitative Dielectric Measurements

Dielectric spectroscopy is a well-known characterization technique that is typically used to analyze macroscopic samples. It provides measurements of dielectric properties in a broad frequency range and at various temperatures. It is quite useful to bring dielectric measurements to the micro and nano scales, and several efforts were already undertaken in this direction. As shown in Eq. (7.22), the electrostatic force at $2\omega_{\mathrm{elec}}$ is related to dC/dZ, with the latter being a direct result of the local dielectric properties. One of the recent studies of local dielectric properties was performed with an AFM tip on the top of a poly(vinyl acetate) film deposited on a conducting substrate [60]. A phase-lock-loop controller was used for the topographic feedback in the FM mode; the electrostatic interactions were stimulated by an AC voltage applied at a much smaller ω_{elec}. The voltage and phase of the signal at $2\omega_{\mathrm{elec}}$ was determined in order to obtain the local dielectric susceptibility. The dependence of the real and imaginary components $V_{2\omega\mathrm{elec}}$ on frequency generally mimics the macroscopic dielectric curves, yet a temperature shift of a few degrees was noticeable between these measurements. The results of the local dielectric measurements at different temperatures also demonstrate the similarity of the frequency responses to those obtained in macroscopic experiments at different temperatures. In addition to measurements of the local dielectric response in one location, the mapping of the dielectric response of a thin film made of PS and PVAC blend was performed at different temperatures [61]. Specifically, it was shown that the domains of PVAC are identified by a strong phase contrast that appears near the glass transition temperature of this polymer. The matrix, which is presumably

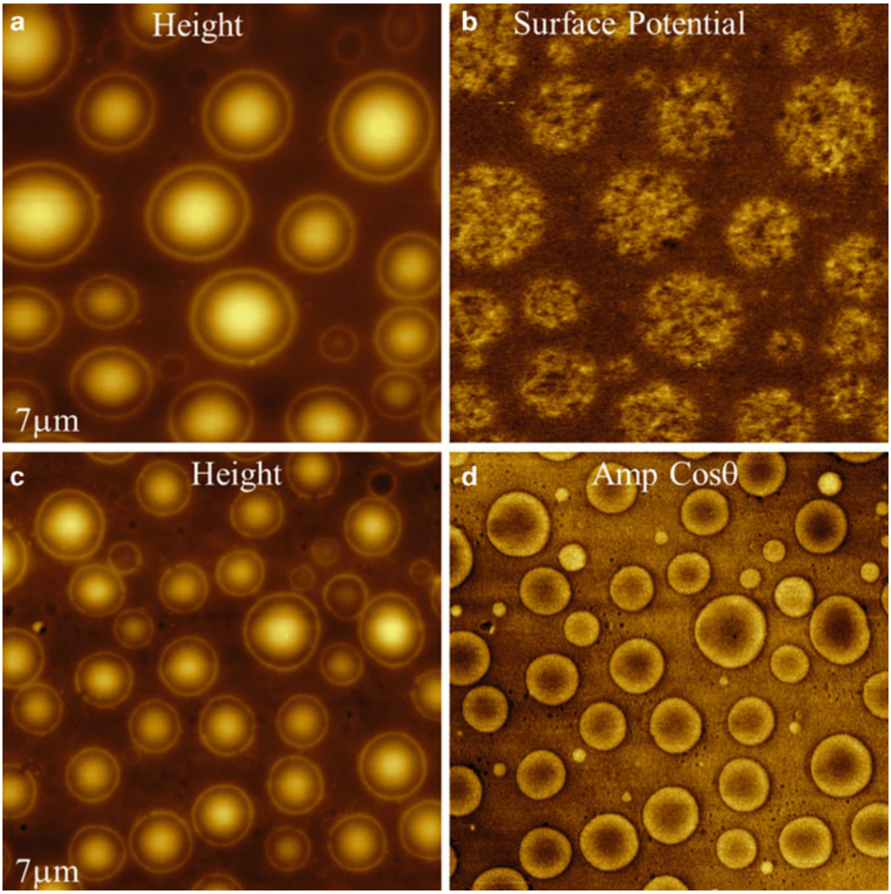


Fig. 7.24 Height (**a**, **c**), surface potential (**b**), and amplitude of cos θ at $2\omega_{\text{elec}}$ (**d**) images of a thin film of PS/PVAC blend

enriched in PS, does not change its contrast because the glass transition of PS is much higher.

It is worth noting that the aforementioned studies were conducted in UHV (ultra high vacuum) in noncontact mode. We have selected the same material, PS/PVAC blend, as the test sample for single-pass studies under ambient conditions. Typical images of the thin film of this blend on a conducting ITO substrate are shown in Fig. 7.24a–d. The film morphology is characterized by spherical domains imbedded into the matrix and it is consistent with the immiscible nature of this blend, which leads to phase separation of the constituents, Fig. 7.24a, c. The surface potential of the domains is higher (~200 mV) than the matrix's potential, and this difference correlates with the fact that the dipole moments of the polymer molecules are quite different (PS—0.3D, PVAC—2.1D), Fig. 7.24b. For dielectric measurements

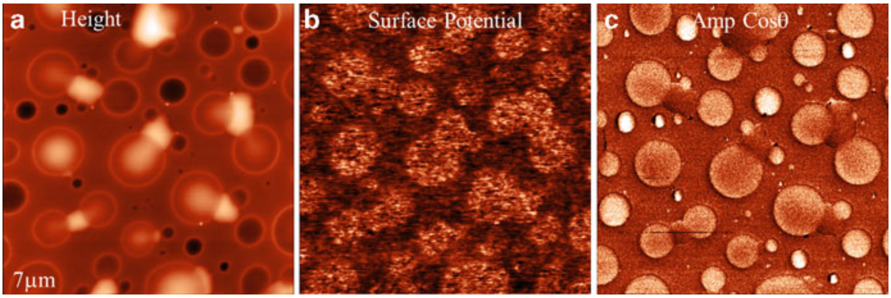


Fig. 7.25 Height (**a**), surface potential (**b**), and amplitude of cos θ at $2\omega_{\text{elec}}$ (**c**) images of a PS/PVAC film, which was annealed at 80 °C

we applied the same scheme as described in Fig. 7.19 but the 2nd LIA was tuned to $2\omega_{\text{elec}}$. In phase modulation experiments typical values for ω_{elec} are predominantly in the 3–5 kHz range. In most dielectric studies we detect the amplitude and phase signals of cos θ at $2\omega_{\text{elec}}$. An amplitude image is shown for PS/PVAC film in Fig. 7.24d. The analysis of this image, which is substantially different from the surface potential image, is quite complicated. At first glance, one can notice that the pattern generally resembles the reversed topography profile of this film.

Annealing of PS/PVAC at temperatures above the glass transition of PVAC (~40 °C) and below the glass transition of PS (~100 °C) induces morphology changes, which most likely reflect the flow of PVAC polymers from elevated domains to nearby surface regions. This leads to the formation of the elevated patches in between the spherical domains enriched in PVAC, Fig. 7.25a. The surface potential image in Fig. 7.25b confirms that the elevated patches are of the same nature as the spherical domains. The contrast of the dielectric response is quite different, with the related pattern (Fig. 7.25c) exhibiting the most pronounced features at the spherical depressions. Again, the map of the amplitude of cos θ at $2\omega_{\text{elec}}$ mimics the inverse topography profile.

The quantitative analysis of the dielectric response is more complicated than in the case of the surface potential, which is directly measured in KFM studies. The nanoscale capacitance of a thin dielectric film depends on the ratio of film thickness and dielectric permittivity [62]. Therefore, the topography-related contribution complicates the images containing local dielectric properties, and this effect should be considered in the analysis. The other problem is related to the existing methods for dielectric studies. High-contrast images related to local dielectric properties were obtained using the response at the 2nd flexural mode. This data is more difficult to treat theoretically because the signal is enhanced through the cantilever resonance and its *Q*-factor should be taken into account. A situation with the analysis of the cos $2\omega_{\text{elec}}$ (amplitude and phase) signal is more straightforward. According to the theoretical description of the probe motion in the oscillatory AFM mode there is a general relationship between the cosine phase and the force acting on a probe. This equation can be applied for electrostatic tip–sample force interactions and the related integral is calculated analytically [63]. As a result, one obtains the relationship between amplitude of phase cosine ($G_{2\omega_{\mathrm{elec}}}^{\cos\theta}$), the capacitance of the tip–sample junction, the probe features, and a ratio of sample thickness to permittivity.

$$\left[G_{2\omega_{\mathrm{elec}}}^{\cos\theta}\right]\frac{2A_0k\tilde{R}}{\pi Q_1U_{\mathrm{ac}}^2\varepsilon_0R\bar{A}}=\frac{2}{\bar{A}^2}\left[\frac{1}{\sqrt{1-\bar{A}^2x^2}}-\frac{1+x}{\sqrt{(1+x)^2-\bar{A}^2x^2}}\right], \qquad (7.27)$$

where ε_0 is the vacuum dielectric constant and ε_{r} is the relative dielectric constant of the film; Q_1 is the quality factor, h the thickness of the film, and Z_{c} the apex-film separation distance; θ_0 is the tip cone angle; R is the effective apex radius; $\tilde{R}=R[1-\sin\theta_0]$; A and A_0 are the actual and free amplitudes $\bar{A}=A/\tilde{R}$; $x=\frac{\tilde{R}}{Z_{\mathrm{c}}+p}$; and $p=\frac{h}{\varepsilon_{\mathrm{r}}}$ [63].

This relationship can be applied for the extraction of quantitative permittivity values from the experimentally measured cosine phase, the capacitance, and the probe parameters. We have used this approach to get quantitative data for two polymer films using a LabVIEW-based program that incorporates the above formulas [59].

A verification of the quantitative dielectric measurements was performed on thin, homogeneous PS and PVAC films, which were prepared by spin-casting solutions of the pure polymers in toluene on conducting ITO glass substrates. A sharp wooden stick was used to scratch through the films so that the imaging of the polymer film thickness becomes possible. The morphology from one location of the PS film on ITO glass is shown in Fig. 7.26a. Simultaneously with height images, the cosine phase response on the polymer film and substrate was detected at different stimulating AC voltages, Fig. 7.26b. A quadratic dependence of the amplitude versus voltage is observed for the applied voltages up to 2 V, as judged by the cross-section profiles taken across the PS and ITO locations, Fig. 7.26c, d. The calculations of the dielectric permittivity were made for measurements performed at a stimulating AC voltage of 1 V (ω_{elec} = 4 kHz) with probes having different tip radii. The results showed that at this frequency the permittivity of PVAC is ~1.4 times higher than that of PS and the averaged absolute values ($\varepsilon_{\mathrm{PS}}$ = 1.7; $\varepsilon_{\mathrm{PVAC}}$ = 2.6) are close to those determined in macroscopic measurements. The local dielectric measurements of these and other polymer films at elevated temperatures and different frequencies are in progress.

7.6 Conclusions

AFM is developing as the multifunctional characterization technique with an emphasis on the measurements in the submicron scale. The AFM studies are advancing in high-resolution imaging and examination of local mechanical and

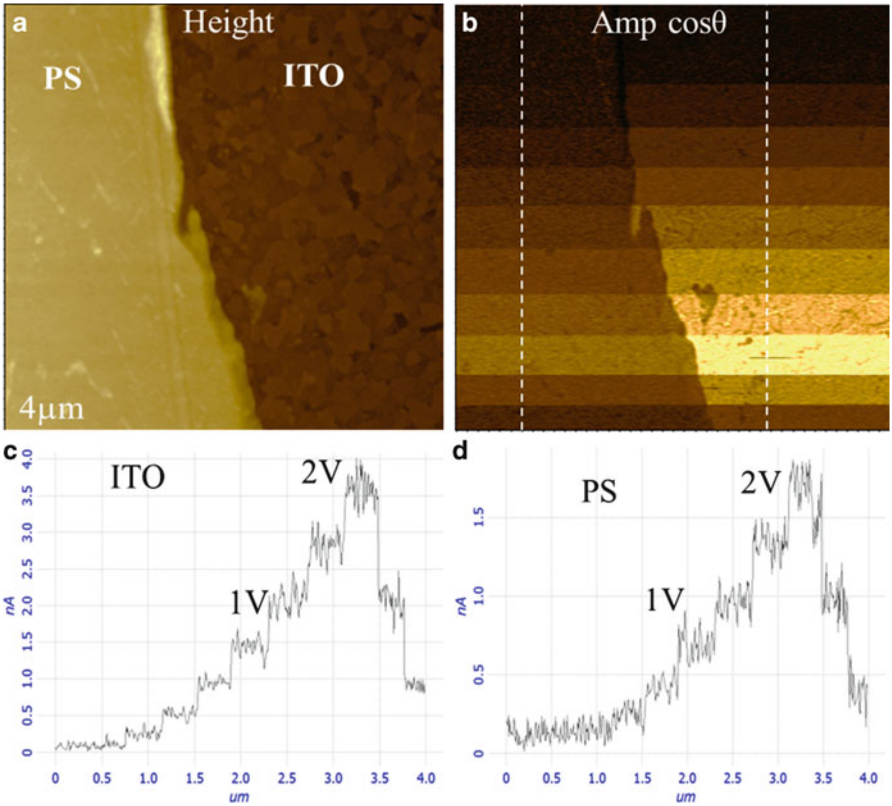


Fig. 7.26 (**a**, **b**) Height and amplitude of cos θ at $2\omega_{\mathrm{elec}}$ images of a scratched location of a PS film on ITO glass. During scanning the stimulated AC voltage was changed. (**c**, **d**) Vertical profiles taken across the polymer and the substrate in the amplitude of cos θ image

electric properties, and we have described the recent efforts in these applications areas. The current progress in instrumentation development is focused on the improvement of the microscope sensitivity and multifrequency approaches. It is also desirable that environmental AFM capabilities will find strong support of the developers. No doubt the instrumental improvements will be accompanied by new results, although, their analysis and understanding are needed to be supported by theoretical efforts. The computer modeling of several important issues of the AFM experiments has been described in detail and the interplay between the theory and practice is invaluable on the way to quantitative studies of mechanical, electric, and dielectric properties at small scales. We would like to emphasize the novel character of local dielectric measurements that were initiated 3–4 years ago. So far, only the initial steps were undertaken in this field and its further development most likely depends on the expansion of such studies to broader frequency and temperature ranges. The complex dielectric permittivity data, which can be extracted from such measurements, will help in revealing molecular motions in nanoscale functional structures. A combination of structural studies with simultaneous detection of the surface potential and complex dielectric permittivity will substantially enhance the role of AFM characterization of photovoltaic cells, batteries, and other small-scale devices.

The described efforts towards the AFM-based quantitative measurements of the mechanical and electric properties pointed out the importance of the probe characterization. This is the area where the joint efforts of AFM practitioners and the probe manufacturers are needed. The availability of a spectrum of well-characterized probes will be of great help to researchers in optimization of the experiments.

The applications areas described in this chapter are the most developed ones. There is room for even better characterization of the materials in the submicron scale by combining AFM with spectral techniques such as Raman scattering and infra-red microscopy. Such alliances can provide the chemical information about surfaces and thus enhance the compositional mapping of heterogeneous materials. The recent results of polymer blends obtained with tip-enhanced studies of a combined AFM-confocal Raman instrument demonstrated that chemical sensitive mapping can be performed at sub-100 nm scale [64].

References

1. G. Binnig, H. Rohrer, C. Gerber, E. Weibel, Surface studies by scanning tunneling microscopy. Phys. Rev. Lett. **49**, 57–61 (1982)
2. R. Young, J. Ward, F. Scire, The topographiner: An instrument for measuring surface microtopography. Rev. Sci. Instrum. **43**, 999–1011 (1972)
3. G. Binnig, C.F. Quate, C. Gerber, Atomic force microscope. Phys. Rev. Lett. **56**, 930–933 (1986)
4. G. Schmalz, Uber Glätte und Ebenheit als physikalisches und physiologishes problem. Z. Vereines Deutscher Ingenieure. **Oct 12**, 1461–1467 (1929)
5. J.E. Sader, S.P. Jarvis, Accurate formulas for interaction force and energy in frequency modulation force spectroscopy. Appl. Phys. Lett. **84**, 1801–1803 (2004)
6. R. Garcia, P. Perez, Dynamic atomic force microscopy methods. Surf. Sci. Rep. **47**, 197–301 (2002)
7. R.W. Stark, W.M. Heckl, Rev. Sci. Instrum. **74**, 5111 (2003)
8. S. Timoshenko, *Vibration Problems in Engineering*, 3rd edn. (D van Nostrand, New York, 1955)
9. N. Krylov, N. Bogolubov, *Introduction to Non-Linear Mechanics* (Princeton University Press, Princeton, NJ, 1949)
10. T. Fukuma, T. Ichii, K. Kobayashi, H. Yamada, K. Matsushige, True-molecular resolution imaging by frequency modulation atomic force microscopy in various environments. Appl. Phys. Lett. **86**, 034103–034105 (1995)
11. D. Maugis, *Contact, Adhesion and Rupture of Elastic Solids* (Springer, Berlin, 2000)
12. S. Belikov, S. Magonov, Classification of dynamic atomic force microscopy control modes based on asymptotic nonlinear mechanics, in Proceedings of American Control Society, St. Louis, pp. 979–985 (2009)

13. S. Belikov, S. Magonov, Tip-sample interaction force modeling for AFM simulation, control design, and material property measurement, in Proceedings of American Control Conference, pp. 2867–2872 (2011)
14. S. Belikov, N. Erina, S. Magonov, Interplay between an experiment and theory in probing mechanical properties and phase imaging of heterogeneous polymer materials. J. Phys. Conf. Ser. **6**, 765–769 (2007)
15. S. Belikov, S. Magonov, True molecular-scale imaging in atomic force microscopy: Experiment and modeling. Jpn. J. Appl. Phys. **45**, 2158–2165 (2006)
16. Q. Zhong, D. Innis, K. Kjoller, V. Elings, Fractured polymer/silica fiber surface studied by tapping mode atomic force microscopy. Surf. Sci. Lett. **290**, L688–L692 (1993)
17. B. Mesa, S. Magonov, Novel diamond/sapphire probes for scanning probe microscopy applications. J. Phys. Conf. Ser. **61**, 770–774 (2007)
18. S.N. Magonov, AFM in Analysis of Polymers, in *Encyclopedia of Analytical Chemistry*, ed. by R.A. Meyers (Wiley, Chichester, 2000), pp. 7432–7491
19. S.N. Magonov, V. Elings, J. Cleveland, D. Denley, M.-H. Whangbo, Tapping-mode atomic force microscopy study of the near-surface composition of a styrene-butadiene-styrene triblock copolymer film. Surf. Sci. **389**, 201–211 (1997)
20. W. Stocker, G. Bar, M. Kunz, M. Möller, S.N. Magonov, H.-J. Cantow, Atomic force Microscopy on polymers and polymer related compounds 2. Monocrystals of normal and cyclic alkanes $C_{33}H_{68}$, $C_{36}H_{74}$, $C_{48}H_{96}$, $C_{72}H_{144}$. Polym. Bull. **26**, 215–222 (1991)
21. C.C. McGonigal, R.H. Bernhardt, D.J. Thomson, Imaging alkane layers at the liquid/graphite interface with the scanning tunneling microscope. Appl. Phys. Lett. **57**, 28–30 (1990)
22. W. Hofbauer, R.J. Ho, R. Hairulnizam, N.N. Gosvami, S.J. O'Shea, Crystalline structure and squeeze-out dissipation of liquid solvation layers observed by small-amplitude dynamic AFM. Phys. Rev. B **80**, 134104–134109 (2009)
23. D. Klinov, S. Magonov, True molecular resolution in tapping mode atomic force microscopy. Appl. Phys. Lett. **84**, 2697–2699 (2004)
24. R.I. Gearba, A.I. Bondar, M. Lehmann, B. Goderis, W. Bras, M.H.J. Koch, D.A. Ivanov, Templating crystal growth at the nano-scale with a thermotropic columnar mesophase. Adv. Mater. **17**, 671–676 (2005)
25. S.M. Lindsay, T. Thundat, L. Nagahara, U. Knipping, R.L. Rill, Images of the DNA double helix in water. Science **244**, 1063–1064 (1989)
26. J. Kumaki, T. Hashimoto, Conformational change in an isolated single synthetic polymer chain on a mica surface observed by atomic force microscopy. J. Am. Chem. Soc. **125**, 4907–4917 (2003)
27. I.S. Yermolenko, V.K. Lishko, T.P. Ugarova, S.N. Magonov, High-resolution visualization of fibrinogen molecules and fibrin fibrils with atomic force microscopy. Biomacromolecules **12**, 370–379 (2011)
28. A.L. Weisenhorn, P.K. Hansma, T.R. Albrecht, C.F. Quate, Forces in atomic force microscopy in air and water. Appl. Phys. Lett. **54**, 2651–2653 (1989)
29. B.V. Derjaguin, The force between the molecules. Sci. Am. **203**, 47–53 (1960)
30. D. Tabor, R.H.S. Winterton, Surface forces: Direct measurement of normal and retarded van der Waals forces. Nature **219**, 1120–1128 (1968)
31. S.I. Bulychev, S.I. Alekhin, M.K. Shorshorov, A.P. Ternovskii, G.D. Shnyrev, Determination of Young's modulus according to the indentation diagram. Ind. Lab. **41**, 1409–1412 (1975)
32. N.A. Burnham, R.J. Colton, Measuring the nanomechanical properties and surface forces of materials using an atomic force microscope. J. Vac. Sci. Technol. A **7**, 2906–2913 (1989)
33. H.-U. Krotil, T. Stifter, H. Waschipky, K. Weishaupt, S. Hild, O. Marti, Pulsed force mode: A new method for the investigation of surface properties. Surf. Interface Anal. **27**, 336–340 (1999)
34. O. Sahin, S. Magonov, C. Su, C.F. Quate, O. Solgaard, An atomic force microscopy tip designed to measure time-varying nanomechanical forces. Nat. Nanotechnol. **2**, 507–514 (2007)

35. A.F. Sarioglu, O. Solgaard, Cantilevers with integrated sensor for time-resolved force measurements in tapping-mode atomic force microscopy. Appl. Phys. Lett. **93**, 023114–023116 (2008)
36. J.L. Hutter, J. Bechhoefer, Calibration of atomic-force microscope tips. Rev. Sci. Instrum. **64**, 1868–1873 (1993)
37. H. Hertz, Uber die Beruhrung fester elasticher Korper. J. Reine Angew. Math. **92**, 156–171 (1882)
38. I. Sneddon, The relation between load and penetration in the axisymmetric Boussinesq problem for a punch of arbitrary profile. Int. J. Eng. Sci. **3**, 47–57 (1965)
39. S. Belikov et al., J. Phys. Conf. Ser. **61**, 1303–1307 (2007)
40. S. Belikov, N. Erina, L. Huang, C. Su, C. Prater, S. Magonov, V. Ginzburg, B. McIntyre, H. Lakrout, G. Meyers, Parametrization of atomic force microscopy tip shape models for quantitative nanomechanical measurements. J. Vac. Sci. Technol. B **27**, 984–992 (2009)
41. W. Oliver, G. Pharr, Measurement of hardness and elastic modulus by instrumented indentation: Advances in understanding and refinements to methodology. J. Mater. Res. **19**, 3–20 (2004)
42. S.N. Magonov, D.H. Reneker, Characterization of polymer surfaces with atomic force microscopy. Annu. Rev. Mater. Sci. **27**, 175–222 (1997)
43. B.V. Derjaguin, V.M. Miller, Y.P. Toporov, Effect of contact deformations on the adhesion of particles. J. Colloid. Interface Sci. **53**, 314–326 (1975)
44. K.L. Johnson, K. Kendall, A.D. Roberts, Surface energy and the contact of elastic solids. Proc. R. Soc. Lond. A **324**, 301–313 (1971)
45. S.N. Magonov, N.A. Yerina, Scanning Probe Microscopy of Elastomers and Rubbery Materials, Chap. 23, ed. by A. Bhowmick. Current Topics of Elastomers Research (Taylor & Francis Group, Oxford, 2008)
46. Y. Martin, D.A. Abraham, H.K. Wickramasinghe, High-resolution capacitance measurement and potentiometry by force microscopy. Appl. Phys. Lett. **52**, 1103–10005 (1988)
47. K. Kobayashi, H. Yamada, K. Matsushige, Dopant profiling on semiconducting sample by scanning capacitance force microscopy. Appl. Phys. Lett. **81**, 2629 (2002)
48. V.B. Elings, J.A. Gurley, Scanning probe microscope using stored data for vertical probe positioning. U.S. Patent 5,308,974 (1994)
49. J. Alexander, S. Magonov, M. Moeller, Topography and surface potential in Kelvin force microscopy of perfluoroalkyl alkanes self-assemblies. J. Vac. Sci. Technol. B **27**, 903–911 (2009)
50. S. Magonov, J. Alexander, Beilstein J. Nanotechnol. **2**, 15–27 (2011)
51. U. Zerweck, C. Loppacher, T. Otto, S. Grafstroem, L.M. Eng, Accuracy and resolution limits of Kelvin probe force microscopy. Phys. Rev. B **71**, 125424–125433 (2005)
52. F. Krok, K. Sajewicz, J. Konior, M. Goryl, P. Piatkowski, M. Szymonski, Lateral resolution and potential sensitivity in Kelvin probe force microscopy: Towards understanding of the sub-nanometer resolution. Phys. Rev. B **77**, 235427–235429 (2008)
53. J. Colchero, A. Gil, A.M. Baro, Resolution enhancement and improved data interpretation in electrostatic force microscopy. Phys. Rev. B **64**, 245403–14 (2001)
54. S. Magonov, J. Alexander, J. Belikov, Single-Pass Measurements in Atomic Force Microscopy: Kelvin Force Microscopy and Local Dielectric Studies. Application Note, January 2012, NT-MDT Development Inc. (2012)
55. F. Giessibl, Forces and frequency shifts in atomic-resolution dynamic-force microscopy. Phys. Rev. B **56**, 16010–16015 (1997)
56. S. Magonov, J. Alexander, S.-H. Jeoung, N. Kotov, High-resolution imaging of molecular and nanoparticles self-assemblies with Kelvin force microscopy. J. Nanosci. Nanotechnol. **10**, 1–5 (2010)
57. A. Mourran, B. Tartsch, M. Gallyamov, S. Magonov, D. Lambreva, B.I. Ostrovskii, I.P. Dolbnya, W.H. de Jeu, M. Moeller, Self-assembly of the perfluoroalkyl-alkane F14H20 in ultrathin films. Langmuir **21**, 2308–2316 (2005)

58. A. El Abed, M.-C. Faure, E. Pouzet, O. Abilon, Experimental evidence for an original two-dimensional phase structure: An antiparallel semifluorinated monolayer at the air-water interface. Phys. Rev. E **5**, 051603–051604 (2002)
59. S. Magonov, J. Alexander, S. Wu, Advancing Characterization of Materials with Atomic Force Microscopy – Based Electric Techniques, in *Scanning Probe Microscopy of Functional Materials: Nanoscale Imaging and Spectroscopy*, ed. by S.V. Kalinin, A. Gruverman (Springer, Berlin, 2010), pp. 233–300
60. P.S. Crider, M.R. Majewski, J. Zhang, H. Oukris, N.E. Israeloff, Local dielectric spectroscopy of near-surface glassy polymer dynamics. J. Chem. Phys. **128**, 044908–5 (2008)
61. C. Riedel, R. Arinero, P. Tordjeman, G. Lévêque, G.A. Schwartz, A. Alegria, J. Colmenero, Nanodielectric mapping of a model polystyrene-poly(vinyl acetate) blend by electrostatic force microscopy. Phys. Rev. E **81**, 010801–010804 (2010)
62. G. Gomila, J. Toset, L. Fumagali, Nanoscale capacitance microscopy of thin dielectric films. J. Appl. Phys. **104**(024315), 1–8 (2008)
63. S. Belilkov, J. Alexander, S. Magonov, I. Yermolenko, Atomic force microscopy control system for electrostatic measurements based on mechanical and electrical modulation. American Control Conference, Montreal 3228–3233 (2012)
64. L. Xue, W. Li, G.G. Hoffmann, J.G.P. Goossens, J. Loos, G. de With, High-resolution chemical identification of polymer blend thin films using tip-enhanced Raman mapping. Macromolecules **44**, 2852–2858 (2011)

Index

A. Korkin and D.J. Lockwood (eds.), *Nanoscale Applications for Information and Energy Systems*, Nanostructure Science and Technology,
DOI 10.1007/978-1-4614-5016-0, © Springer Science+Business Media New York 2013