Mohammad D. Al-Amri
Mohamed M. El-Gomati
M. Suhail Zubairy
Editors

Optics in Our Time

OPEN

Optics in Our Time

Mohammad D. Al-Amri
Mohamed M. El-Gomati
M. Suhail Zubairy

Optics in Our Time

Springer Open

Editors
Mohammad D. Al-Amri
National Center for Applied Physics
King Abdulaziz City for Science and Technology
Riyadh, Saudi Arabia

Mohamed M. El-Gomati
Department of Electronics
University of York
Heslington, UK

M. Suhail Zubairy
Department of Physics and Astronomy
Texas A&M University
College Station, TX, USA

ISBN 978-3-319-31902-5 ISBN 978-3-319-31903-2 (eBook)
DOI 10.1007/978-3-319-31903-2

Library of Congress Control Number: 2016959602

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Light has occupied a position of importance in our attempt to understand the world around us. The earliest studies going back to the dawn of civilization related to our attempts to understand vision and the properties of optical materials. The modern era in optics is rooted in the great work of Ibn al-Haytham whose work on the nature of light and its applications had a long-lasting impact. In our generation, the discovery of laser has opened up not only new areas of research but also had great impact on a number of technologies. Lasers have revolutionized the fields of communication, medicine, and biotechnology. It has influenced the art, architecture, and printing. It is therefore befitting that United Nations has declared 2015 as the International Year of Light and Light-Based Technologies to celebrate these great achievements. Saudi Arabia is one of the sponsors of this initiative that has ignited a number of activities all around the world. This volume, which covers the history of light and its applications to many diverse branches of science, contains articles by some of the leading scientists who have played a key role in advancing the frontiers in our own times.

Turki S.M. Al-Saud
King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

Preface

Light and light-based technologies have played an important role in transforming our lives via scientific contributions spanned over thousands of years. In this book, we present a vast collection of articles on various aspects of light and its applications in the contemporary world at a popular or semi-popular level. These chapters are written by the world authorities in their respective fields. This is therefore a rare volume where the world experts have come together to present the developments in this most important field of science in an almost pedagogical manner.

This volume covers five aspects related to light. The first presents two articles, one on the history of the nature of light and the other on the scientific achievements of Ibn al-Haitham (Alhazen), who is broadly considered the father of modern optics. These are then followed by an article on ultrafast phenomena and the invisible world. The third part includes papers on specific sources of light, the discoveries of which have revolutionized optical technologies in our lifetime. They discuss the nature and the characteristics of lasers, solid-state lighting based on the light emitting diode (LED) technology, and finally modern electron optics and its relationship to the Muslim golden age in science. The book's fourth part discusses various applications of optics and light in today's world, including biophotonics, art, optical communication, nanotechnology, the eye as an optical instrument, remote sensing, and optics in medicine. In turn, the last part focuses on quantum optics, a modern field that grew out of the interaction of light and matter. Topics addressed include atom optics, slow, stored and stationary light, optical tests of the foundation of physics, quantum mechanical properties of light fields carrying orbital angular momentum, quantum communication, and wave-particle dualism in action.

We are grateful to many individuals and organizations whose contributions and cooperation were invaluable in compiling this book. First and foremost, we are grateful to all the authors who took their time in writing these articles for the general audience. We are very grateful to the leadership and the staff at *King Abdulaziz City for Science and Technology* (KACST) for their generous support in the completion of this project. Khalid Al Zahrani ought to be thanked with whom the idea of this book was triggered over a cup of tea.

We have however one deep regret: one of the authors, Nobel Laureate Ahmed Zewail, who enthusiastically supported this volume and contributed an important chapter passed away on August 2, 2016, before the publication of this book.

Mohammad D. Al-Amri
Riyadh, Saudi Arabia

Mohamed M. El-Gomati
York, UK

M. Suhail Zubairy
College Station, TX, USA

Contents

V Quantum Optics

About the Editors and Authors

Mohammad D. Al-Amri

received his MSc in physics with distinction from Sussex University, UK, in 2001. He went on and got his PhD from York University and was the recipient of the Stott Prize in Physics for the best PhD thesis in 2004. He then joined the National Center for Applied Physics (NCAP) at King Abdulaziz City for Science and Technology, where he has been working as a professor.

He was the recipient of the CO/ICTP Gallieno Denardo Award Winner (2013) and the senior membership of the Optical Society of America (2012). He has been working on different areas of research related to quantum optics and quantum informatics, where the focus is on quantum optical lithography and microscopy, weak measurement, and direct quantum communication. He has published around 55 journal papers, within about 35 published in the Journal of PRL and PRA, and has got 5 patents. Some of Dr. AlAmri's research work has been highlighted in semipopular press. He has given many lectures, seminars, and invited talks at universities and international conferences.

Mohamed M. El-Gomati

is professor of electronics at the University of York, UK. His research interests are in the fields of surface science and electron optics with particular emphasis on the development of novel instrumentation for nanostructure analysis. He is the author and co-author of more than 200 articles and patents in these fields. He is a fellow of the Institute of Physics (IoP) and the Royal Microscopical Society (RMS). His interests extend to history of physics with particular emphasis on history of optics within Muslim civilization. He is the chairman of the Foundation for Science, Technology and Civilisation (UK) and a trustee of the educational charity, Curriculum Enrichment for the Future, and is an advisor to a number of UK and overseas universities. He was awarded the UKESCA Award (1993), the Cosslett Award by the Microbeam Society of America (2008), the Fazlur Rahman Prize for Science and Engineering (2009), and the British Muslims Award for Science (2013). In 2012, Professor El-Gomati was awarded an OBE for his services to science.

M. Suhail Zubairy

is a university distinguished professor of physics and the holder of the Munnerlyn-Heep Chair in Quantum Optics at the Texas A&M University. He received his PhD from the University of Rochester in 1978. He served as professor of electronics and the founding chairman of the Department of Electronics at the Quaid-i-Azam University before joining Texas A&M University in 2000. Prof. Zubairy's research interests include quantum optics and laser physics. He has published over 300 research papers on topics such as precision microscopy and lithography, quantum computing, noise-free amplification, and atomic coherence effects. He is the co-author of two books, one on quantum optics and the other on quantum computing devices. He has received many honors including the Willis E. Lamb Award for Laser Science and Quantum Optics, Alexander von Humboldt Research Prize, the Outstanding Physicist Award from the Organization of Islamic Countries, the Abdus Salam Prize in Physics, the International Khwarizmi Award from the president of Iran, the Orders of Hilal-e-Imtiaz and Sitara-e-Imtiaz from the president of Pakistan, and the George H. W. Bush Award for Excellence in International Research. He is an elected member of the Pakistan Academy of Sciences and a fellow of the American Physical Society and the Optical Society of America.

Govind P. Agrawal

received the MS and PhD degrees from the Indian Institute of Technology, New Delhi, in 1971 and 1974, respectively. After holding positions at the Ecole Polytechnique, France, the City University of New York, and AT&T Bell Laboratories, Dr. Agrawal joined in 1989 the faculty of the Institute of Optics at University of Rochester, where he is currently James C. Wyant Professor of Optics. His research interests focus on optical communications, nonlinear photonics, and laser physics. He is an author or co-author of more than 400 research papers and eight books. His books on nonlinear fiber optics (Academic Press, 5th ed., 2013) and fiber-optic communication systems (Wiley, 4th ed., 2010) are used worldwide for research and teaching. Since 2014, he is serving as editor-in-chief of the journal *Advances in Optics and Photonics*.

Prof. Agrawal is a fellow of the IEEE and OSA (the Optical Society) and a life fellow of the Optical Society of India. In 2012, the IEEE Photonics Society honored him with its prestigious Quantum Electronics Award. He received in 2013 the Riker University Award for Excellence in Graduate Teaching. More recently, he was awarded the 2015 Esther Hoffman Beller Medal of the Optical Society.

Pablo Artal

is a full professor of optics at the University of Murcia, Spain. He spent several periods doing collaborative research in laboratories in Europe, Australia, and the USA. He is a fellow member of the OSA, ARVO, and EOS. He received the prestigious 2013 Edwin H. Land Medal Award, and he is the recipient of the exclusive "ERC advanced grant" in 2013. He received the "Rey Jaime I" Award for Applied Research in 2015. He has published more than 180 reviewed papers that received 7000 citations (h-index: 43) and presented more than 150 invited talks in international meetings and is also a co-inventor of 20 international patents. He has pioneered a number of highly innovative advances in the methods for studying the optics of the eye and has contributed substantially to our understanding of the factors that limit human visual resolution. Dr. Artal is the founder of Voptica SL, a spin-off company developing the concept he invented of adaptive optics vision analyzers. He has been the mentor of many graduate and postdoctoral students. His personal science blog is followed by readers, mostly graduate students and fellow researchers, from around the world. He has been editor of the *Journal of the Optical Society of America A* and the *Journal of Vision*.

Robert W. Boyd

was born in Buffalo, New York. He received the BS degree in physics from MIT and the PhD degree in physics from the University of California at Berkeley. His PhD thesis was supervised by Charles Townes and involves the use of nonlinear optical techniques in infrared detection for astronomy. Professor Boyd joined the faculty of the University of Rochester in 1977 and in 2001 became the M. Parker Givens Professor of Optics and Professor of Physics. In 2010, he became professor of physics and Canada Excellence Research Chair in Quantum Nonlinear Optics at the University of Ottawa. His research interests include studies of "slow" and "fast" light propagation, quantum imaging techniques, nonlinear optical interactions, studies of the nonlinear optical properties of materials, and the development of photonic devices including photonic biosensors. Professor Boyd has written two books, co-edited two anthologies, published over 400 research papers (≈29,000 citations, Google h-index 71), and been awarded nine patents. He is the 2009 recipient of the Willis E. Lamb Award for Laser Science and Quantum Optics, the 2010 recipient of a Humboldt Research Prize, and the 2014 recipient of the Quantum Electronics Award of the IEEE Photonics Society.

Charles M. Falco

has joint appointments as professor of optical sciences and professor of physics at the University of Arizona where he holds the UA Chair of Condensed Matter Physics. He is a fellow of four professional societies (the American Physical Society, the Institute of Electrical and Electronics Engineers (IEEE), the Optical Society of America, and the Society of Photo-optical Instrumentation Engineers (SPIE)), has published more than 275 scientific manuscripts, co-edited two books, has seven US patents, and has given over 400 invited talks at conferences, research institutions, and cultural organizations in 33 countries. In addition to his scientific research, he was co-curator of the Solomon R. Guggenheim museum's "The Art of the Motorcycle" which, with over 2 million visitors in New York, Chicago, Bilbao, and the Guggenheim Las Vegas, was by far the most successful exhibition of industrial design ever assembled. More recently, he and the world-renowned artist David Hockney found artists of such repute as van Eyck, Bellini, and Caravaggio who used optical projections in creating portions of their work. Three international conferences have been organized around these discoveries, and recognition for them includes the 2008 Ziegfeld Lecture Award from the National Art Education Association.

Michael Fleischhauer

is professor of theoretical physics at the University of Kaiserslautern, Germany. He made his PhD in physics at the University of Friedrich-Schiller University Jena on the theory of nonclassical light and his Habilitation on Electromagnetically Induced Transparency (EIT) and its applications at the Ludwig Maximilian University (LMU) of Munich, both in Germany. His interests are in quantum optics and many-body physics of ultracold quantum gases. He is an expert in numerical methods for strongly interacting systems in low dimensions, and his research interests include topological systems. Among other things, he has developed the method of light storage using EIT, which is by now one of the main techniques for quantum memories for light and key for photon-based quantum networks. He served as department head of the selection committee of the Alexander von Humboldt Foundation and at the boards of several physics journals and is currently a member of the editorial board of *Physical Review Letters*. Michael Fleischhauer is member of the executive board of the German Physical Society where he has been the spokesperson of the division of quantum optics and photonics. He is a member of the senate of the German Research Foundation and has been elected to the Academy of Sciences and Literature in Mainz.

Edward S. Fry

distinguished professor and former head of the Department of Physics and Astronomy at Texas A&M University, holds the George P. Mitchell Chair in Experimental Physics and has been with Texas A&M since 1969. He is past chair of the Texas Section of the American Physical Society and is a fellow of both the American Physical Society and the Optical Society of America. He received the Association of Former Students Distinguished Faculty Teaching Award (1993), the Texas A&M Distinguished Scientist Award of Sigma Xi (2001), and the Association of Former Students Distinguished Faculty Achievement Award (2012).

Dr. Fry's research interests cover the gamut from basic research to applied research. Some notable achievements include (i) one of the first, and definitive, Bell inequality tests of the foundations of quantum mechanics, addressing questions first raised by Einstein; (ii) the first observations of lasing without (population) inversion (LWI); (iii) a new integrating cavity technique for the measurement of optical absorption in the presence of even severe scattering, leading to what are now widely considered the standard reference data for pure water absorption; and (iv) a new diffuse reflector whose reflectivity is so high that ring-down spectroscopy in an integrating cavity is now possible.

Colin J. Humphreys

is professor of materials science and director of research in the Department of Materials Science and Metallurgy, University of Cambridge, and a fellow of Selwyn College, Cambridge. He is a fellow of the Royal Society and a fellow of the Royal Academy of Engineering. He founded and directs the Cambridge Centre for Gallium Nitride (GaN). He founded two spin-off companies to exploit the research of his group on low-cost LEDs for home and office lighting. The companies were acquired in February 2012 by Plessey, which is now manufacturing LEDs based on this technology at their factory in Plymouth, UK. He also founded and directs the Cambridge/Rolls-Royce Centre for Advanced Materials for Aerospace.

Gediminas Juzeliūnas

is a director of the Institute of Theoretical Physics and Astronomy of Vilnius University, Lithuania. He is also a principal researcher (research professor) at the institute. Dr. Juzeliunas completed a PhD in 1986 in theoretical condensed matter physics at Vilnius University, studying optical properties of excitons in confined geometries. Subsequently he held a two-year postdoctoral appointment at the University of East Anglia, England, shifting his research area toward quantum optics. Dr. Juzeliunas was a Humboldt research fellow at the University of Ulm, Germany (1997–1998), and a Fulbright scholar at the University of Oregon in the USA (2000–2001). Dr. Juzeliunas received a National State Prize for Science of Lithuania in 2008, a Vilnius University Rector's Award in 2010, and a Jucys Prize for Theoretical Physics of the Lithuanian Academy of Science in 2013. His current research focuses on ultracold atomic gases, slow light, and metamaterials. In particular, this includes a pioneering theoretical work on light-induced gauge potential for ultracold atoms.

Olga Kocharovskaya

is the distinguished professor in the Department of Physics and Astronomy, Texas A&M University. She joined the Texas A&M faculty in 1998 after 12 years at the Institute of Applied Physics of the Russian Academy of Sciences. She made a number of contributions to laser science and quantum optics.

These include the predictions of the phenomena of electromagnetically induced transparency and lasing without inversion as well as suggestion and experimental realization of the various schemes for coherent control of gamma-ray nuclear transitions. A fellow of both the American Physical Society and Optical Society of America, she has earned the Willis E. Lamb Medal for Laser Physics and Quantum Electronics, the Sigma Xi Distinguished Scientist Award, the Texas A&M Association of Former Students Distinguished Achievement Award in Research, and the Texas A&M University Distinguished Professor Award.

Mario Krenn

is a PhD student since 2012 in the group of Anton Zeilinger at the University of Vienna and Institute for Quantum Optics and Quantum Information (IQOQI) of the Austrian Academy of Sciences. He is investigating quantum properties of spatial structures of photons and high-dimensional Hilbert spaces. This involves the certification and long-distance transmission of spatial mode entanglement. In connection with complex high-dimensional multipartite quantum states, he is interested in automated computer-designed quantum experiments. In his free time, he enjoys playing table soccer and fantasizing about artificial general intelligence.

Mehul Malik

is a Marie Curie postdoctoral fellow in the group of Professor Anton Zeilinger at the Institute for Quantum Optics and Quantum Information in Vienna, Austria. Originally from New Delhi, India, Mehul received his PhD in optics in 2013 from the University of Rochester under the supervision of Professor Robert Boyd and a bachelor of arts in 2006 from Colgate University. He is currently working on creating the first multi-photon entangled states in high dimensions using the orbital angular momentum property of light. His broader research interests lie in the fields of fundamental quantum optics, quantum imaging, and quantum information. Mehul has published 19 papers in internationally recognized journals such as *Reviews of Modern Physics*, *Nature Communications*, and *Physical Review Letters*. He has delivered talks on his research in 11 different countries, including 5 invited talks at international conferences. Outside the laboratory, Mehul is an avid skier, loves to cook, and enjoys dancing salsa.

Alexis Méndez

received a PhD degree in electrical engineering from Brown University, USA, in 1992. He is president of MCH Engineering LLC—a consulting firm specializing in optical fiber sensing technology. Dr. Mendez was the former group leader of the Fiber Optic Sensors Lab within ABB Corporate Research (USA) where he led R&D activities for the development of fiber sensors for use in industrial plant, oil and gas, and high-voltage electric power applications. He has written 60 technical publications, taught several short courses on fiber sensors, holds 5 US patents, and is a recipient of an R&D100 award.

Dr. Mendez is a member of the OFS International Steering Committee, is a fellow of SPIE, and was past chairman of the 2006 International Optical Fiber Sensors Conference (OFS-18) and past technical chair of the 2nd Workshop on Specialty Optical Fibers and Their Applications (WSOF-2). He is also a member of the International Society for Health Monitoring of Intelligent Infrastructure (ISHMII) Committee. He is co-editor of the *Specialty Optical Fibers Handbook* and co-author of SPIE's *Fiber Optical Sensors Book: Fundamentals and Applications, 4th Ed.*

Pierre Meystre

obtained his physics diploma and PhD from the Swiss Federal Institute of Technology in Lausanne, Switzerland, and the Habilitation in Theoretical Physics from the University of Munich, Germany. He is a Regents Professor of Physics and Optical Sciences at the University of Arizona and since 2013 serves as lead editor of *Physical Review Letters*. His research interests include theoretical quantum optics, atomic physics, ultracold science, and quantum optomechanics. He has published well over 300 refereed papers and is the author of the text *Elements of Quantum Optics*, together with Murray Sargent III, and of the monograph "Atom Optics." He is a recipient of the Humboldt Foundation Research Prize for Senior US Scientists, the R.W. Wood Prize of the Optical Society of America, and the Willis E. Lamb Award for Laser Science and Quantum Optics. He is a fellow of the American Physical Society, the Optical Society of America, and the American Association for the Advancement of Science.

Munir H. Nayfeh

nanotechnology pioneer, Islamic Academy of Science fellow, and physics professor at the University of Illinois, received his BSc from the American University of Beirut and PhD from Stanford University in atomic laser spectroscopy. He was a researcher at Oak Ridge National Laboratory, lecturer at Yale University, and consultant at Argonne National Laboratory. He is chairman of the advisory board at center of excellence in nanotechnology (King Fahd University), ex-vice-chairman of the International Science Council at King Abdullah Institute for Nanotechnology, and advisor at the Nanotechnology Center at An-Najah University and Saigon Hi-Tech Park, Vietnam. He served on panels for King Abdullah University (KAUST) and

Royal Commission, Jubail. Nayfeh co-authored *Electricity and Magnetism* (translated into Farsi) and co-edited three laser books. He presents nanotechnology science fiction using the trademark "Dr. Nano." He holds the largest number of patents in nanosilicon worldwide (22 US-European, 18 issued) and is a founder of nanotechnology companies: NanoSi Advanced Technologies, Nano Silicon Solar, and Parasat-Nanosi (Kazakhstan). Professor Nayfeh developed a process for creating highly luminescent ultrasmall silicon nanoparticles with electronics, photonics, and biomedicine applications, a process for combining lasers with high electric fields of scanning electron microscopes to pin/write atoms on surfaces with atomic resolution, and a process using strong laser light to detect single atoms.

Roshdi Rashed

is emeritus research director (distinguished class) at the Centre National de la Recherche Scientifique (CNRS, Paris) and the former director of the Center for History of Arabic and Medieval Sciences and Philosophy at Paris VII—Denis Diderot University. He is honorary professor at the University of Tokyo and emeritus professor at the University of Mansoura. From 1965 on, his field of research was the history of mathematics and mathematical sciences and their applications. He is the author of about fifty books, mainly in French and in English, including six substantial volumes devoted to Ibn al-Haytham's mathematics, optics, and astronomy. These volumes, written in French—with a critical edition of Ibn al-Haytham's texts—have been translated into English and Arabic.

Bahaa Saleh

has been dean of CREOL, the College of Optics and Photonics at the University of Central Florida, since 2009. He was born in Cairo, Egypt, and received the PhD degree from Johns Hopkins University in 1971. He held positions at the University of Santa Catarina in Brazil; Kuwait University; Max Planck Institute in Germany; University of California-Berkeley; European Molecular Biology Laboratory; Columbia University; University of Wisconsin-Madison, where he was chair of electrical and computer engineering (1990–1994); and Boston University, where he was chair of electrical and computer engineering (1994–2008). He has made significant contributions to coherence and statistical optics, nonlinear optics, quantum optics, and image science, and his publications include more than 600 journal and conference papers and 3 books: *Photoelectron Statistics* (Springer, 1978), *Fundamentals of Photonics* (Wiley, 2007, with M. C. Teich), and *Subsurface Imaging* (Cambridge, 2011). He served as editor-in-chief of the *Journal of the Optical Society of America A* (1991–1997) and founding editor of OSA's *Advances in Optics and Photonics* (2008–2013). He is a fellow of the IEEE, OSA, SPIE, APS, and Guggenheim Foundation and a recipient of the OSA Beller Award, OSA Meese Medal, OSA Distinguished Service Award, SPIE BACUS award, and Kuwait Prize.

Thomas Scheidl

is a senior scientist at the Institute for Quantum Optics and Quantum Information of the Austrian Academy of Sciences and a team leader within the group of Prof. Anton Zeilinger. He studied experimental physics at the University of Vienna and received a doctor's degree in 2009. His main field of research is the experimental investigation of fundamental questions in quantum physics as well as the development of prototypes for secure quantum communications and quantum information, mainly for free-space systems. He was experimentally active in many international collaborations and responsible for a number of projects funded by the European Space Agency. He has published more than 12 papers in scientific journals such as in *Nature*, *Nature Physics*, and *Physical Review Letters*, and in 2012 and 2013, he received fellowships from the Energie AG (Innovation Fund) and from the Sohmen Far East Foundation for his contribution to the field of long-distance quantum communication.

Wolfgang P. Schleich

is engaged in research on quantum optics ranging from the foundations of quantum physics via general relativity to number theory. He was educated at the Ludwig Maximilian University (LMU) of Munich and studied with Marlan O. Scully at the University of New Mexico, Albuquerque, and the Max Planck Institute for Quantum Optics, Garching. Moreover, he was also a postdoctoral fellow with John Archibald Wheeler at the University of Texas at Austin. Professor Schleich is a member of several national and international academies and has received numerous prizes and honors for his scientific work such as the Gottfried Wilhelm Leibniz Prize, the Max Planck Research Award, and the Willis E. Lamb Award for Laser Science and Quantum Optics. He is also a distinguished adjunct professor at the University of North Texas and a faculty fellow at Texas A&M University Institute for Advanced Study. His textbook, *Quantum Optics in Phase Space*, has been translated into Russian, and a Chinese edition was published in 2010.

Marlan O. Scully

(Baylor, Princeton, and Texas A&M) has worked on a variety of problems in laser physics and quantum optics including the first quantum theory of the laser with Lamb, the laser phase transition analogy and its applications to the Bose condensate, experimental demonstrations of lasing without inversion, and ultraslow light in hot gases via quantum coherence. His introduction of entanglement interferometry to quantum optics has shed light on the foundations of quantum mechanics, e.g., the quantum eraser. Recently, he and his colleagues have applied quantum coherence to remote sensing of anthrax and probing through turbid medium such as skin and plant tissue. Scully is currently a distinguished university professor at Texas A&M University and also holds positions at Princeton and Baylor Universities. He has been elected to the US National Academy of Sciences and the Max Planck Society. He has recently been awarded the OSA Frederic Ives Medal/Quinn Prize, the DPG/OSA Herbert Walther Award, and the Commemorative Silver Medal of the Senate of the Czech Republic (by K. Chapin).

Yanhua H. Shih

professor of physics at the University of Maryland, Baltimore Campus (UMBC), received his PhD in 1987 from the University of Maryland at College Park, USA. He started the Quantum Optics Laboratory at UMBC in the fall of 1989. His laboratory has been recognized as one of the leading groups in the field of quantum optics that attempts to probe the foundations of quantum theory. In the past 10 years, he published more than 100 papers in leading refereed journals and given more than 100 invited presentations in national and international professional conferences and workshops. His book *An Introduction to Quantum Optics: Photon and Biphoton Physics*, published in 2011, is a good summary of his theoretical and experimental research. Yanhua Shih is a winner of the 2002 Willis E. Lamb Medal for pioneering contributions to quantum electronics and especially the study of spatial coherence effects of multi-photon entangled states.

Alexei V. Sokolov

obtained a master's diploma from Moscow Institute of Physics and Technology (1994) and a physics PhD from Stanford University (2001). Currently at Texas A&M University, Sokolov holds a professor position in physics and astronomy and a Stephen Harris Professorship in Quantum Optics. His overall expertise is in the field of laser physics, nonlinear optics, ultrafast science, and spectroscopy. His research interests center around applications of molecular coherence to quantum optics, ultrafast laser science and technology including generation of sub-cycle optical pulses with prescribed temporal shapes, and studies of ultrafast atomic, molecular, and nuclear processes, as well as applications of quantum coherence in biological and defense-oriented areas. Sokolov is an OSA fellow; his awards include the Lomb Medal (OSA, 2003), the Hyer Award (TX section APS, 2007), and the Treat Award (Texas A&M Research Foundation, 2011).

Rupert Ursin

is a group leader and senior scientist at the IQOQI. His main field of research is to develop quantum communication and quantum information processing technologies, mainly for free-space but also for fiber-based systems. The scope of his work ranges from near-term engineering solutions for secure key sharing (quantum cryptography) to more speculative research (de-coherence of entangled states in gravitational fields). Experiments on quantum communication and teleportation using entangled photon pairs are also among his interests, with the long-term goal of a future global quantum network based on quantum repeaters. He has published more than 49 papers in peer-reviewed scientific journals such as *Science and Nature*. He has been experimentally active in numerous international collaborations in Germany, Italy, Spain, and the USA, as well as in Japan. To date, several of his publications were selected as yearly highlights by the British PhysicsWeb and others. In 2008, he received the Award for the Telecommunications Advancement Research Fellowship (National Institute of Information and Communications Technology (NICT), Tokyo, Japan) and in 2010, the Christian Doppler Prize. He has presented invited talks on original scientific results at more than 60 prestigious international conferences. He currently holds a guest professorship at the University of Science and Technology (USTC) in Shanghai, China.

Dmitri V. Voronine

is a research assistant professor at the Institute for Quantum Science and Engineering, Texas A&M University, and a visiting scholar at Baylor University. His research interests are in the experimental and theoretical spectroscopy, quantum optics, and ultrafast nano-photonics with applications to the investigation, imaging, and control of complex systems. He has made contributions in coherent control of nano-optical excitations and multidimensional spectroscopy, time-resolved surface-enhanced coherent Raman spectroscopy, and quantum biophotonics.

Thomas Walther

studied physics at the Ludwig Maximilian University (LMU) of Munich. After receiving his PhD from the University of Zürich in physical chemistry in 1994, he joined the Physics Department of Texas A&M University, College Station, TX, as a research assistant scientist, where he became an assistant professor in 1998. In 2002, he moved to the TU Darmstadt, Germany, as a full professor. His research interests include basic research in quantum optics as well as lasers and their applications.

Vladislav V. Yakovlev

is currently professor in the Department of Biomedical Engineering at Texas A&M University. He also holds a joint appointment in the Department of Physics and Astronomy. He got his BS/MS/PhD degrees in physics/quantum electronics from Moscow State University, Russia. He completed his postdoctoral training in the Department of Chemistry and Biochemistry, University of California-San Diego, where he first developed the concept of optical parametric amplification of white-light continuum, which is nowadays employed as a standard tool in ultrafast optical spectroscopy, and then performed a series of groundbreaking experiments demonstrating coherent control of molecular systems. Since joining the Physics Department at the University of Wisconsin-Milwaukee, Dr. Yakovlev has been involved in instrumentation development for biomedical research. He was the first to demonstrate broadband coherent anti-Stokes Raman microscopy, Raman photoacoustic imaging, and deep-tissue coherent Raman imaging, including anthrax detection in the mail. In 2012, he moved to Texas A&M University where he continuously developed new spectroscopic tools for biological, medical, and environmental sensing and imaging. He has published over 200 research publications and gave over 150 scientific talks. Dr. Yakovlev is a fellow of the Optical Society of America, the American Institute of Medical and Biological Engineering, and the International Society for

Optics and Photonics. His current research interests include biomechanics on a microscale level, nanoscopic optical imaging of molecular and cellular structures, protein spectroscopy and structural dynamics, bioanalytical applications of optical technology and spectroscopy, environmental sensing, and deep-tissue imaging and sensing.

Anton Zeilinger

(PhD 1971, Vienna University) is professor of physics at the University of Vienna and president of the Austrian Academy of Sciences. He has performed many groundbreaking experiments in quantum mechanics, from important fundamental tests all the way to innovative applications. Most of his research concerns the fundamental aspects and applications of quantum entanglement. He is one of the pioneers of the new field of quantum information science. The most important stations of his career include M.I.T., the Collége de France, the Technical Universities of Munich and Vienna, Oxford University, and the University of Innsbruck. Among his many awards and prizes are the German Order Pour le Mérite, Wolf Prize, the Inaugural Isaac Newton Medal of the Institute of Physics, the Medaille du Collége de France, the Great Cross of Merit with Star of the Federal Republic of Germany, and the King Faisal Prize. He is a member of the Austrian, Berlin-Brandenburg, Slovak, Belarus, and Serbian Academies of Science, the German National Academy Leopoldina, the Académie des Sciences Paris, the US National Academy of Sciences, and the European Academy of Sciences and a fellow of the American Association for the Advancement of Science (AAAS), the American Physical Society (APS), and the World Academy of Sciences (TWAS).

Ahmed H. Zewail

is the Linus Pauling Chair Professor of Chemistry and Physics and director of the Center for Physical Biology at Caltech. He is the sole recipient of the 1999 Nobel Prize for the development of the field of *femtochemistry*. In the post-Nobel era, he developed *4D electron microscopy* for the direct visualization of matter in space and time. Dr. Zewail's other honors include fifty honorary degrees, orders of merits, postage stamps, and more than a hundred international awards.

He has published some 600 articles and 14 books and is known for his effective public lectures and writings, not only on science but also in global affairs. For his leadership role in these world affairs, he received, among others, the "Top American Leaders Award" from *The Washington Post* and Harvard University. In 2009, President Barack Obama appointed him to the Council of Advisors on Science and Technology, and in the same year, he was named the first US Science Envoy to the Middle East. Subsequently, the Secretary General of the United Nations Ban Ki-moon invited Dr. Zewail to join the UN Scientific Advisory Board. In Egypt, he serves in the Council of Advisors to the President. Following the 2011 Egyptian revolution, the government established "Zewail City of Science and Technology" as the national project for scientific renaissance, and Dr. Zewail became its first chairman of the board of trustees.

Aleksey Zheltikov

received his PhD degree from M.V. Lomonosov Moscow State University in 1990. He received his doctor of science degree from the same university in 1999. He has been a professor at M.V. Lomonosov Moscow State University since 2000, the head of Neurophotonics Laboratory at Kurchatov Institute Russian Research Center since 2010, and the head of Advanced Photonics International Laboratory at the Russian Quantum Center since 2012. Since 2010, he is a professor at Texas A&M University. He is the author of more than 600 peer-reviewed scientific publications. He is the winner of the Russian Federation State Prize for young researchers (1997), Lamb Award for achievements in quantum electronics (2010), Shuvalov Prize for research at Moscow State University (2001), and Kurchatov Prize for achievements in neurophotonics (2014).

Dandan Zhu

received the BE and MS degrees in materials science from Tsinghua University, Beijing, China, in 2001 and 2003, respectively, and the PhD degree in materials science from the University of Cambridge, UK, in 2007.

Her PhD project was on metal-organic vapor phase epitaxy growth and characterization of near-UV emitting quantum well and light-emitting diode (LED) structures on sapphire substrates. After graduation, she worked as a research associate with the Cambridge Centre for Gallium Nitride, focusing on the development of crack-free LED structures on large area Si (111) substrates. She subsequently joined Plessey Semiconductors as chief materials engineer, working on technology transfer and R&D on GaN-on-Si technology for mass production of LEDs. In 2015, she returned to academia to continue her research interest in nitride materials. She has over 30 publications and 2 patents in the field of nitride-based LEDs.

History

Contents

A *Very* Brief History of Light

M. Suhail Zubairy

M.S. Zubairy (✉)
Institute for Quantum Science and Engineering (IQSE) and Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843-4242, USA
e-mail: zubairy@tamu.edu

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_1

1

1.1 Introduction

While tracing the history of ideas that shaped our understanding of nature and the properties of light, it is quite remarkable to see how one can almost neatly divide the geographical regions where human thoughts progressed during a certain time period followed by a decay and setting of the dark ages. We can divide the history of light into four distinct eras. The first era, with its center initially in Athens and then Alexandria, belonged to the Greeks. This era extended from about 800 BC till around 200 AD. It seems that hardly anything of significance in our understanding of light was contributed between 200 AD till around 750 AD when Muslims burst onto the scene. The second era belongs to the Islamic civilization, with its centers in Baghdad and Cordoba. It had its golden age till around middle to late thirteenth century when Mongol invasion destroyed the eastern center in Baghdad in 1258 and the decay set in the Western Center of Cordoba. The third era started in Europe around the fourteenth century when medieval Europe that had slipped into a dark age after the fall of Roman Empire started to emerge out of it. The crusades (1095–1272) and the conquest of Islamic Spain made the Muslim scholarship and the Greek traditions accessible to the Europeans, helping to initiate the glorious era of scientific revolution in the West. The last era started with the dawn of twentieth century that opened not only with new and revolutionary theories of Physics but also with a revolution in communication technology. This has helped to make science, and optics, a global preoccupation.

1.2 Greeks and Antiquity

The Greek civilization flourished in the eastern Mediterranean area, extending from Athens in Greece to Anatolia, Syria, and Egypt from Archaic period in about eighth century BC till about 200 AD. This civilization produced the highest level of intellect in many branches of human thought such as mathematics, philosophy, ethics, and astronomy. Through the galaxy of thinkers, such as Archimedes, Socrates, Plato, Aristotle, Euclid, Ptolemy, and Galen, they left a lasting imprint on the human civilization. Their most lasting legacy is not the theories that these giants of history presented as most of them have either been overturned or replaced during the evolution of human thought. Their lasting legacy to the mankind lies in placing the rational thinking at the apex of creation that has reverberated through millennia, long after the Greek civilization disappeared.

Light in Antiquity The earliest studies concerning light had to do with understanding vision. For example, the ancient Egyptians believed that light was the activity of their god Ra seeing. When Ra's eye (the Sun) was open, it was day. When it was closed, night fell. The earliest studies on the nature of light and vision can be attributed to the Greek and Hellenistic traditions. The Greek period, extending from the Archaic period till around 320 BC and centered in Athens, produced many earliest ideas about vision through the works of Democritus, Epicurus, Plato, and Aristotle. After the death of Alexander, the center shifted to Alexandria where Ptolemy I, a general in the Alexander's army, established a new dynasty that lasted till the Roman conquest of Egypt in the first century BC. In this Hellenistic period, the glorious traditions of Greek scholarship in the field of light and vision continued through the works of Euclid, Hero of Alexandria, Ptolemy, and Galen.

The theory of vision attempts to explain how objects, near and far, their shape, size, and color, are perceived by us. The earliest systematic studies of vision are

attributed to atomists who reduced every sensation, including vision, to the impact of atoms from the observed object on the organ of observation. There were different schools of thought among atomists. For example, Democritus (460 BC–370 BC) believed that the visual image did not arise directly in the eye, but the air between the object and the eye is contracted and stamped by the object seen and the observing eye. The pressed air contains the details of the object and this information is transferred to the eye. Epicurus (341 BC–270 BC), on the other hand, proposed that atoms flow continuously from the body of the object into the eye. However the body does not shrink because other particles replace and fill in the empty space.

An alternate theory of vision due to Plato (428 BC–328 BC) and his followers advocated that light consisted of rays emitted by the eyes. The striking of the rays on the object allows the viewer to perceive things such as the color, shape, and size of the object. Our vision was initiated by our eyes reaching out to "touch" or feel something at a distance. This is the essence of extramission theory of light that would be influential for almost a 1000 years until Alhazen would conclusively prove it to be wrong.

Hellenistic Era, Euclid, Hero of Alexandria, and Ptolemy Euclid (b. 300 BC) is the father of Geometry. His book *Elements* laid down the foundation of axiomatic approach to geometry and is one of the most influential books ever written. Little original references are available about Euclid and what we know about him was written centuries after he lived by Proclus (c. 450 AD) and Pappus of Alexandria (c. 320 AD). His work in optics follows the same methodology as *Elements* and gives a geometrical treatment of the subject. Euclid believed in extramission and his theory of vision is founded in the following postulates:

1. Rectilinear rays proceeding from the eye diverge indefinitely;
2. The figure contained by the set of visual rays is a cone of which the vertex is at the eye and the base at the surface of the object seen;
3. Those things are seen upon which visual rays fall and those things are not seen upon which visual rays do not fall;
4. Things seen under a larger angle appear larger, those under a smaller angle appear smaller, and those under equal angles appear equal;
5. Things seen by higher visual rays appear higher, and things seen by lower visual rays appear lower;
6. Similarly, things seen by rays further to the right appear further to the right, and things seen by the rays further to the left appear further to the left;
7. Things seen under more angles are seen more clearly.

Euclid did not define the physical nature of these visual rays. However, using the principles of geometry, he discussed the effects of perspective and the rounding of things seen at a distance.

Euclid had restricted his analysis to vision. Hero of Alexandria (10–70), who also believed in the extramission theory of Euclid, extended the principles of geometrical optics to consider the problems of catoptrics, particularly, reflection from smooth surfaces. Hero derived the law of reflection by invoking the principle of least distance. According to him, light from a point A to another point B follows a path that is shortest. On this basis, he showed that when light reflects from a surface, angle of incidence is equal to the angle of reflection. Specifically, the image appears to be as far behind the mirror as the object is in front of the mirror. Hero's principle of least distance would be replaced by the principle of least time by Pierre Fermat more than 1500 years later to derive the law of refraction.

The most influential and perhaps last important figure in optics of the Greek–Egyptian era was Claudius Ptolemy (90–168). He is most well known for

championing the geocentric model for the movement of planets, a view that would survive for almost 1400 years until it was replaced by a heliocentric model through the work of Nicholas Copernicus in 1543. His book on the subject *Almagest* was very influential in shaping the thinking on astronomy and, along with *Elements* by Euclid, was the longest read book in the history of science. Ptolemy wrote *Optics* in which he discussed the theory of vision, reflection, refraction, and optical illusions. Like Euclid and Hero, Ptolemy championed the extramission theory of vision. He considered visual rays as propagating from the eye to the object seen. However, instead of considering visual rays as discrete lines as postulated by Euclid, he considered them forming a continuous cone. Ptolemy carried out careful experiments on refraction and concluded that, for light propagating from one medium to another, the ratio of the angle of incidence to the angle of refraction was constant and depended on the properties of the two media. He thus derived the small angle approximation of the law of refraction. The formulation of theory based on experimental results, frequently supported by the construction of special apparatus, is the most striking feature of Ptolemy's *Optics*.

1.3 Islamic Period

Islam has its roots in Mekkah, a city that was on the cross road of trade route from Syria to Yemen in the sixth and seventh century. The founder of the religion, Prophet Muhammad (PBUH), was born there in 570 and claimed to have received his first revelations from God in 610. Under severe opposition from his kinsmen to the new religion, he migrated to the northern city of Madinah in 622. This marked the beginning of the Islamic era. When Muhammad (PBUH) died in 632, Islam had spread throughout the Arabian Peninsula. He was followed by four caliphs, Abu Bakar, Umar, Usman, and Ali, in the leadership of the Islamic community. Under Umar, the conquests of Persia, Syria, and Egypt expanded the Islamic writ to a major part of the Middle East. These caliphs were followed by the Ummayad dynasty (660–750) when North Africa, Spain, Western China (Xinjiang), and Western India (modern Pakistan) came under the Muslim rule. The capital of Ummayads was Damascus. In 750, the Ummayads were replaced by Abbasids (descendants of an uncle of the Prophet named Abbas). They continued to rule till 1258 when Mongols attacked and conquered their capital Baghdad. The Ummayad's rule in the Iberian Peninsula continued till 1492 (the year when Columbus landed in the new world). The tenth and eleventh century Egypt was ruled by Fatimids, a dynasty founded by the descendants of Fatima, daughter of Prophet Muhammad (PBUH).

Contrary to some modern claims and perceptions, Islam was an enlightened religion in its beginning, deeply rooted in the search of knowledge. The Islamic holy book, Qur'an, that was believed by Muslims to be the direct word of God as revealed to Prophet Muhammad (PBUH) exhorted human beings to contemplate and seek knowledge through words such as "And say, Lord increase my knowledge" (Qur'an 20:114) and "He (God) has subjected to you, as from Him, all that is in the heavens and on earth: behold in that are Signs indeed for those who reflect" (Qur'an 45:13). Similarly the sayings, such as "the ink of a scholar is more holy than the blood of a martyr" and "The most learned of men is the one who gathers knowledge from others on his own; the most worthy of men is the most knowing and the meanest is the most ignorant", attributed to Prophet Muhammad (PBUH) emphasize the importance of the pursuit of knowledge. These and other similar injunctions in the Qur'an and the Prophetic traditions helped to develop an attitude in the Muslim community that supported the quest of knowledge and promoted an environment where open discussion was encouraged. Science was not seen to be contrary to the faith. Rather it was considered to be a religious duty

to seek knowledge and understanding. As the Islamic Empire increased in size so did the thirst for more knowledge in all fields.

Armed with this attitude, Muslims built a civilization in the Eighth century that would last for several centuries and contributed to almost all aspects of human knowledge. It is unprecedented in the annals of history that an empire would bring with it a great civilization as well. The Muslims built a body of knowledge by first learning and then expanding on the older traditions, particularly Greek. Their own contributions would, in turn, provide a foundation for the emergence of the modern Western civilization.

Bayt-al-Hikmah (House of Wisdom) Traditionally the beginning of the Islamic Golden Age of science is attributed to the Abbasid caliph Harun al-Rashid (763–809) who ruled an empire stretching from modern Pakistan to North Africa to the shores of Atlantic Ocean from 786 till 809. However the age may have started earlier, even in the Ummayad period, when the foundations of Islamic Jurisprudence were being laid down through discussion and reason. This tradition crossed into the secular body of knowledge, leading first to assimilating what old sages from Greek, Indian, and other civilizations had contributed and then building their own contributions in fields ranging from philosophy and medicine to mathematics and physical sciences.

Harun al-Rashid and his court is fantasized in the book *One Thousand and One Nights*. He laid the foundation of *Bayt-al-Hikma* (House of Wisdom) in the newly built capital city of Baghdad. However it was formally completed in 830 in the era of his equally brilliant son, al-Mamun (786–833) who ruled from 813 till 833. Originally the House of Wisdom was a scientific academy and a public library where books from all parts of the empire were brought and translated in Arabic. These included old texts from India, Greece, and Persia in the fields of philosophy, mathematics, astronomy, medicine, and optics. By 850, House of Wisdom had the largest repository of manuscripts of its time. Gradually this center turned into a center of research and many famous names of Islamic Golden era were associated with it. These included, among others, Jabir bin Hayyan (721–815) who is regarded as father of chemistry, Al-Khwarizmi (780–850) who is credited with inventing algebra, Al-Kindi (800–873) who is regarded as the first Muslim philosopher, Hunayn ibn-Ishaq (809–873) whose contributions in medicine were influential till the modern era, and Alhazen (865–1040) who is regarded as the father of optics.

Al-Mamun himself took great interest in the progress of the House of Wisdom and is reputed to have intellectual discussions with the scholars that had started coming from distant lands. Al-Mamun supported organized research in areas such as developing detailed maps of the world, measurement of the circumference of the Earth, and the confirmation of data from Ptolemy's *Almagest*. This is the first known example of the state sponsored research. Gradually other institutions of higher learning, such as Al-Azhar University (970) in Cairo and Al-Nizamiyya (1095) in Baghdad, developed in and outside Baghdad.

Al-Kindi and Optics Abu Yusef Yaqoub ibn Ishaq Al-Kindi (800–873) was the first great philosopher of the Islamic era. He synthesized, adopted, and promoted Greek philosophy in the Islamic world. He worked with a group of translators at the House of Wisdom who rendered works of Aristotle, Plato, Euclid, and other Greek mathematicians and scientists into Arabic. Al-Kindi's main authority in philosophical matters was Aristotle. His philosophical treatises include *On First Philosophy*, in which he argues that the world is not eternal and that God is a simple *One*. Al-Kindī tried to demonstrate that philosophy is compatible with Islamic traditions and had a great

influence on later Muslim philosophers Abdullah ibn-Sina (known in the West as Avicenna) and Ibn-Rushd (known in the West as Averroes).

Al-Kindi was also the first to undertake serious studies in optics and the theory of vision. His work on optics, *De Aspectibus* in Latin translation, exerted a strong impact on Islamic and Western optics throughout the middle ages. In optics, Al-Kindi followed the traditions of Euclid, and carried on by Ptolemy and others in which geometrical constructions were used to explain phenomena such as vision, reflection, refraction, shadows, and burning mirrors. Whereas Euclid considered the straightness of a visual ray as an axiom, Al-Kindi proved it experimentally by considering the shadows projected by different opaque objects. He treated the geometry of visual cone, rejecting the discreteness of Euclid's rays and replacing them with a cone of continuous beam of radiation similar to Ptolemy.

Al-Kindi's work was followed in tenth century by Al-Razi and Al-Farbi who started objecting to the extramission theory of light. A series of strong arguments against the notion of visual fire were put forward around 1000 by the great ibn-Sina. He argued that the visual fire cannot reach remote objects as it will have to fill an enormous space each time we opened our eyes. He also argued that Euclid's discrete rays may leave large areas of a distant object unobservable.

Ibn-Sahl and Snell's Law 600 Years Before Snell Snell's law of refraction is an important law relating to the propagation of light between two media with different refractive indices. Refractive index of a medium is inversely proportional to the speed of light in that medium. The law of refraction thus forms the basis of understanding the bending of light rays from various kinds of lenses. The credit of the discovery of the law of refraction is given to Willebrord Snellius (1580–1626) who derived it using trigonometric methods in 1621. However recent studies indicate that this law was discovered more than 600 hundred years earlier during the Islamic Golden Age of Baghdad by a scientist named Abu Sad Al Alla Ibn Sahl (940–1000). Ibn Sahl excelled in optics and wrote a treatise *On Burning Mirrors and Lenses* in 984 in which he discussed the focusing properties of the parabolic and elliptical burning mirrors. He also presented an analysis of how hyperbolic glass lenses bend and focus light. As a lemma in his derivation of the focusing property of light by a plano-convex hyperbolic lens, he presented a geometric argument based on the sine law of refraction. This appears to be a major achievement and shows how far Muslims had advanced in pure and applied mathematics as well as optics by the end of tenth century. A question, however, remains about how such a major discovery could remain ignored for so many centuries. A plausible explanation is that Ibn Sahl did not state the law of refraction explicitly. Instead it was hidden as a sort of lemma and his emphasis was on the focusing property of lenses.

Alhazen, Father of Modern Optics Abu Ali al-Hasan ibn al- Hasan ibn al-Haytham (965–1040), known in the west as Alhazen, is a central figure in science. He is often described as the greatest physicist between Archimedes and Newton. He was the first person to follow the scientific method, the systematic observation of physical phenomena and their relation to theory, thus earning the title First Scientist from many. His most important contribution in optics is his book *Kitab-al Manzir* (Book of Optics) which was completed around 1027. This book, comprising seven volumes, was the first comprehensive treatment of optics and covered subjects such as the nature of light, the physiological treatment of eye, and the bending and focusing

properties of lenses and mirrors. This book was most influential in the transition from the Greek ideas about light and vision to the modern day optics. Alhazen's Book of Optics was translated in Latin at the end of twelfth century under the title *De Aspectibus* and would remain the most influential book in optics till Newton's *Opticks* published in 1704.

Alhazen proved the long held theory of Euclid, Hero, and Ptolemy that light originated from the eye to be wrong and showed that light originated from the light sources. He did this by carrying out a simple experiment in a dark room where light was sent through a hole by two lanterns held at different heights outside the room. He could then see two spots on the wall corresponding to the light rays that originated from each lantern passing through the hole onto the wall. When he covered one lantern, the bright spot corresponding to that lantern disappeared. He thus concluded that light does not emanate from the human eye, but is emitted by objects such as lanterns and travels from these objects in straight lines. Based on these experiments, he invented the first pinhole camera (that Kepler would use and call *camera obscura* in the seventeenth century) and explained why the image in a pinhole camera was upside down.

Alhazen's theory of vision was not limited to the description of light rays originating from the objects and entering the eye. He also understood that an explanation of vision must also take into account the anatomical and psychological factors. He proved that the perception of an image occurs not in the eyes but in the brain and that the location of an image is largely determined by psychological factors.

Alhazen did not invent the telescope but he explained how a lens worked as a magnifier. He contended that magnification was due to the bending, or refraction, of light rays at the glass-to-air boundary and not, as was thought, to something in the glass. He correctly deduced that the curvature of the glass, or lens, produced the magnification. He concluded that the magnification takes place at the surface of the lens, and not within it.

His work on catoptrics in Book V of the Book of Optics dealt with problems of reflection from spherical and parabolic mirrors. It also contains a discussion of what is now known as Alhazen's problem. The problem was first formulated by Ptolemy in 150 AD. Draw lines from two points in the plane of a circle such that they meet at a point on the circumference, making equal angles with the normal at that point. The problem was to locate this point. This problem is equivalent to the Billiard table problem: On a circular table there are two balls; at what point along the circumference must one be aimed at in order for it to strike the other after rebounding off the edge. Alhazen's interest in this problem stemmed from the following formulation of the problem: When light is sent from a source towards a spherical mirror, find the point on the mirror where the light will be reflected to the eye of an observer. The problem is insoluble using a compass and a ruler because the solution requires solution of an equation of fourth degree. Alhazen solved this problem geometrically by the aid of a hyperbola intersecting a circle. This problem remained unsolved using algebraic methods for almost a thousand years until it was finally solved in 1997 by the Oxford mathematician Peter M. Neumann.

1.4 Scientific Revolution

The publication in 1543 of Nicholaus Copernicus's *De Revolutionibus Orbium Coelestium* (On the Revolutions of the Heavenly Spheres) marks the beginning of the scientific revolution. He proposed a heliocentric model of the solar system, a system in which Sun was held at rest and all the planets including Earth circled

around it, replacing the long held Ptolemaic geocentric model in which earth was at rest. Without the benefit of the knowledge of the law of gravitation, it was hard to believe how earth could be moving around the sun still maintaining the stability of all objects including the humans on its surface. The hostility to a model that took away the centrality of earth in a solar system was so great (particularly in the Church) that Copernicus could not publish his heliocentric theory till the end of his life. According to a legend, Copernicus received the copy of his book *De Revolutionibus* on the very last day of his life, thus dying without knowing that his work heralded a new era of human history. There were, however, birth pangs of this new world of science, most famous being Galileo's heresy conviction in 1633 for his support of Copernicus.

Kepler The most important figure to follow Copernicus was the German astronomer, Johannes Kepler (1571–1630), whose laws on planetary motion would prove pivotal for Newton's law of gravitation. Kepler is a key figure in the history of light and vision as well. His interest in the subject appears to have originated in his observation of a solar eclipse on July 10, 1600 by means of *camera obscura*. Several years ago, Tycho Brahe (1546–1601), the greatest naked-eye astronomer of the time, had observed that the angular diameter of the moon appeared to be larger during a solar eclipse when observed through the pin-hole camera than when observed directly. Kepler understood that this anomaly could not be explained without a full understanding of the optical instruments, in this case, the *camera obscura*. He noted that the finite diameter of the pinhole should be responsible for this anomaly. He discovered the solution by an experimental technique where he stretched a thread through an aperture from a simulated luminous source to the surface on which the image was formed. He traced out the image cast by each point on the luminous body seeing, in the process, the geometry of radiation in three-dimensional terms. In this way, Kepler was able to formulate a satisfactory theory of radiation through apertures based on the rectilinear propagation of light rays.

Kepler did not stop at explaining Tycho Brahe's problem of seemingly variable lunar diameter. In 1601, he noted that eye itself possesses an aperture and should be treated in the same way as the aperture in a pinhole camera. Kepler published his theory of vision in 1604.

Descartes Until Kepler, the main motivation of studying the nature of light came from a desire to understand vision. René Descartes (1590–1650) appears to be the first person to concern himself with the intrinsic nature of light and the laws of optics. Descartes was a French philosopher and mathematician who had a great impact on western philosophy. He is heralded as the Father of Modern Philosophy. His mathematical contributions included a connection between geometry and algebra that allowed for the solving of geometrical problems using algebraic equations. Descartes promoted the accounting of physical phenomena by way of mechanical explanations.

Descartes main contribution to optics is his book *Dioptrics* that was published in 1637. It deals with many topics relating to the nature of light and the laws of optics. He compares light to a stick that allows a blind person to discern his environment through touch.

Descartes used a tennis ball analogy to derive the laws of reflection and refraction of light. The credit of the discovery of the law of refraction is given to Willebrord Snellius (1580–1626) who derived it using trigonometric methods in

1621. However Snell did not publish his work in his lifetime. Descartes published the law of refraction 16 years after Snell's death, as Descarte's law of refraction.

Fermat and Principle of Least Time Together with Descartes, Pierre de Fermat (1601–1675) was one of the two greatest French mathematicians of the first half of the seventeenth century. A lawyer by profession, Fermat made a number of important contributions in analytical geometry, probability, and number theory. He is most well known for the Fermat's Last Theorem (no three positive integers a, b, c can satisfy the relation $a^n + b^n = c^n$ for any integer n that is larger than 2) that he conjectured in 1621 but could not be proved till 1994.

Fermat's major contribution in optics relates to his derivation of Snell's law using the principle of least time. Just as Hero of Alexandria had derived the law of reflection on the basis of the principle of least distance 1400 years ago, Fermat argued that light rays going from a point located in a region where it propagates with a particular speed to a point in another region where it propagates with a different speed, it would follow a path that takes the shortest time. This yielded the correct Snell's law.

Newton Sir Isaac Newton (1642–1727) is definitely the defining figure in the history of science. His *Principia* laid down the foundation of classical mechanics. His law of gravitation is a bright example of the nature of scientific law–law that applies equally well for all objects, big and small. His contributions in mathematics, particularly his co-discovery of calculus with Wilhelm Leibniz, provided tools that would be so vital for almost all the subsequent major discoveries in physics and other branches of science. They played key role in shaping the physics of the coming centuries.

It is, however, interesting to note that the most important experimental contributions to physics made by Newton are all in the field of optics. He was the first to show that the color is the property of light and not of the medium. Through ingenious experiments he could show that the light generated by sun consisted of all the colors. For example, when sun light passes through a prism, it is dispersed in a rainbow of colors. The red color bends the least and the violet color bends the most. This ability of glass prisms to generate multiple colors was known since antiquity but it was not attributed to light. Instead color was considered as a characteristic of the material. What Newton showed was that when a particular color passed through the prism, no such dispersion took place. In a relatively complicated setup, when these colors were combined together and passed through the prism again, Newton recovered white light, proving that white light consisted of all the colors.

The other major contribution of Newton towards optics is his design of reflecting telescope. All the telescopes through his time were unwieldy refracting telescopes that suffered from chromatic aberrations. The earliest refracting telescope, built in 1608, is credited to Hans Lippershey who got the patent for the design. These refracting telescopes consisted of a convex objective lens and a concave eyepiece. Galileo used this design in 1609. In 1611, Kepler described how a telescope could be made with a convex objective lens and a convex eyepiece lens. Newton designed a reflecting telescope where incoming light is reflected by a concave mirror onto a plane mirror that reflected the light to the observer. This design was simple and less susceptible to chromatic aberrations. All the major telescopes that exist today are improved versions of Newton's reflecting telescope.

Newton was also concerned with the nature of light and advocated corpuscular theory of light. According to him, light is made up of extremely small corpuscles, whereas ordinary matter was made of grosser corpuscles. He speculated that

1

through a kind of alchemical transmutation they change into one another. According to him, "Are not gross Bodies and Light convertible into one another,.. . .. And may not Bodies receive much of their Activity from the Particles of Light which enter their Composition?" It is surprising that Newton advocated corpuscular theory of light when there was evidence that supported the wave behavior. For example, Francesco Grimaldi (1618–1663) made the first observation of the phenomenon that he called diffraction of light. He showed through experimentation that when light passed through a hole, it did not follow a rectilinear path as would be expected if it consisted of particles but took on the shape of a cone. Newton explained that the phenomenon of diffraction was only a special case of refraction that was caused somehow by the ethereal atmosphere near the surface of the bodies. Newton could explain the phenomenon of reflection with his theory. However he could explain refraction by incorrectly assuming that light accelerated upon entering a denser medium because the gravitational pull was stronger.

Huygens, Wave Nature of Light When Newton was expounding a corpuscular nature of light, his contemporary, Christian Huygens (1629–1695) suggested a wave picture of light. Huygens published his results in his *Traite de la lumie*re (Treatise on light) in 1690. Crucial to his wave theory was the result recently obtained by Olaus Romer (1679) that the speed of light is finite. He considered light waves propagating through the *ether* just as sound waves propagate through air. He explained the high but finite speed of light by the elastic collisions of a succession of spheres that made the *ether*. The light waves, according to Huygens, were thus longitudinal waves as opposed to the later studies by Fresnel and Maxwell that showed light to consist of transverse waves.

Huygens formulated a principle (that now bears his name) which describes wave propagation as the interference of secondary wavelets arising from point sources on the existing wavefront. In propagation each ether particle collides with all the surrounding particles so that ". . . around each particle there is made a wave of which that particle is the center."

Young and Double-Slit Experiment Till the beginning of nineteenth century, Newton's status was so great particularly in the British Isles that few dared to challenge his corpuscular theory of light. It was, however, Thomas Young (1773–1829) who, in 1802, conclusively demonstrated the wave nature of light through his double-slit experiment. He described his experiment in these words in *The Course of Lectures on Natural Philosophy and the Mechanical Arts* (1807): ". . .. when a beam of homogeneous light falls on a screen in which there are two very small holes or slits, which may be considered as centres of divergence, from whence the light is diffracted in every direction. In this case, when the two newly formed beams are received on a surface placed so as to intercept them, their light is divided by dark stripes into portions nearly equal, but becoming wider as the surface is more remote from the apertures, so as to subtend very nearly equal angles from the apertures at all distances, and wider also in the same proportion as the apertures are closer to each other. The middle of the two portions is always light, and the bright stripes on each side are at such distances, that the light coming to them from one of the apertures, must have passed through a longer space than that which comes from the other, by an interval which is equal to the breadth of one, two, three, or more of the supposed undulations, while the intervening dark spaces correspond to a difference of half a supposed

undulation, of one and a half, of two and a half, or more." With this he firmly established the wave nature of light.

By repeating his experiment, Young could relate color to wavelength and was able to calculate approximately the wavelengths of the seven colors recognized by Newton that composed white light. According to him ". . .. it appears that the breadth of the undulations constituting the extreme red light must be supposed to be, in air, about one 36 thousandth of an inch, and those of the extreme violet about one 60 thousandth."

The Young's double-slit experiment was not only decisive in debunking Newton's corpuscular theory of light, but it also continued to play a crucial role in our understanding of the nature of light and matter even in the twentieth century. For example, in 2002, *Physics World* published the results of a survey on the all-time Ten Most Beautiful Experiments in Physics. Young's double-slit experiment made not one but two appearances on this prestigious list—at number 1 was the double-slit experiment applied to the interference of electrons and, at number 5 was the original experiment by Young.

Young's double-slit experiment was, however, regarded highly controversial and counterintuitive in his own time. How can a screen uniformly illuminated by a single aperture develop dark fringes with the introduction of a second aperture? And how could the addition of *more* light result in *less* illumination? Young's theory would eventually find broad acceptance, particularly through the works of Fresnel in France.

Fresnel, Theory of Diffraction, and Polarization of Light Augustin Jean Fresnel (1788–1827), a contemporary of Young, championed the wave nature of light based on his work on diffraction. Fresnel began by undertaking experiments with diffraction. He noted that when light passed through a diffractor, one could see a series of dark and bright bands behind the diffractor. However when he blocked one edge of the diffractor, the bright bands within the shadow vanished and the bright bands remained only on the unblocked side of the diffractor. From this, he concluded that the bright bands in the shadow were produced by light coming from both edges and the bright bands, when one edge was blocked, resulted from the reflection of light from one edge of the diffractor. He was able to develop a mathematical theory for these observations based on a wave theory of light and could predict the position of bright and dark lines based on where the vibrations were in phase and out of phase. He published his first paper on wave theory of diffraction in 1815.

An episode indicates the stunning success of the wave nature of light as formulated by Fresnel. In 1819, Fresnel presented his work on wave theory of diffraction in a competition by the French Academy of Sciences. The committee of judges, headed by Francois Arago, included Jean-Baptiste Biot, Pierre-Simon Laplace, and Simeon-Denis Poisson. They were all prominent advocates of Newton's corpuscular theory and were not well disposed to the wave theory of light. Poisson was, however, impressed by Fresnel's submission and extended his calculations to come up with an interesting consequence: "Let parallel light impinge on an opaque disk, the surrounding being perfectly transparent. The disk casts a shadow - of course - but the very center of the shadow will be bright. Succinctly, there is no darkness anywhere along the central perpendicular behind an opaque disk (except immediately behind the disk)". According to the corpuscular theory, there could be no bright spot behind the disk. As Chair of the Committee, Arago asked Fresnel to verify Poisson's prediction and amazingly Fresnel found the bright spot as predicted. This discovery was an impressive vindication of the wave theory and Fresnel won the competition. This spot is now known as "Poisson spot."

Despite the triumph of the wave theory of light, the properties of the polarized light still provided a strong argument in favor of the corpuscular theory, since no explanation from a wave theory had ever been made. Following the success of the wave theory in explaining the interference and diffraction phenomena, Fresnel and Arago embarked upon explaining the properties of the polarized light based on Fresnel's theory. In 1817, Fresnel became the first person to obtain what was later called circularly polarized light. The only hypothesis that could explain the experimental results was that light is a transverse wave. In 1821, Fresnel published a paper in which he claimed that light is a transverse wave. Young had independently reached the same conclusion. The assertion that light is a transverse wave was not readily accepted by many, including Arago. Again Fresnel was vindicated when he could explain the double refraction from the transverse wave hypothesis. This helped to seal the status of light as a transverse wave.

Maxwell and Electromagnetic Waves It was left to James Clerk Maxwell (1831–1879) to complete the classical picture of light as consisting of electric and magnetic waves. This was a truly remarkable outcome of his efforts to unify the two known forces of nature: electric force and magnetic force. It was known through the work of Michael Faraday that a time rate of change of magnetic field yielded electric force. The insight due to Maxwell was that if electricity and magnetism were the two sides of the same coin then a change of electric field should similarly result in a magnetic field. This motivated him to add a term in the Ampere's law that corresponded to a time rate of change of the electric field. This addition immediately yielded a wave equation for an electromagnetic wave propagating at the same velocity as known for light, 3×10^8 m/s. The picture of light that emerged was thus that of undulations of mutually perpendicular electric and magnetic fields propagating. The direction of propagation was perpendicular to both the electric and magnetic fields. Maxwell's results were published in 1865. Thus the light waves were shown to be transverse waves in line with Young and Fresnel as opposed to the picture adopted by Huygens where light was seen as a longitudinal wave propagating through the medium *ether*. This description of light as an electromagnetic wave was experimentally demonstrated by Heinrich Hertz (1857–1894) in 1888.

1.5 Light in Twentieth Century

According to a quote, attributed to Lord Kelvin in an address to the British Association for the Advancement of Science in 1900, "There is nothing new to be discovered in physics now. All that remains is more and more precise measurement." The classical theories of mechanics, electromagnetics, thermodynamic, and, of course, light were firmly in place and it was a justified feeling to believe that the basic laws of nature were fully understood.

There were, however, two "clouds" on the horizon of physics at the dawn of the twentieth century. Interestingly enough, both of these involved light. The first cloud, the Rayleigh–Jeans ultraviolet (UV) catastrophe and the nature of blackbody radiation, led to the advent of quantum mechanics, which of course was a radical change in physical thought up to that point. The second cloud, namely the null result of the Michelson–Morley experiment, led to special relativity, which is the epitome of classical mechanics, and the logical capstone of classical physics. These theories, quantum mechanics and the theory of relativity, were major departures from the classical theory as first formally introduced by Newton.

They would shape the physics of the twentieth century. They also dramatically revised our understanding of the nature of light.

Black-Body Radiation, Kirchhoff and Planck The concept of a black body was introduced by Gustav Kirchhoff (1824–1887) in 1860. Kirchhoff knew from looking at the spectral lines from the sun that there was heat energy in empty space, and postulated equilibrium radiation. However the knowledge of what it consisted of was still primitive. By 1860, Maxwell's equations had not yet been postulated, and the electromagnetic nature of both heat and light rays had not yet been established. Nor had the existence of atoms in the walls of a cavity, nor that an oscillator radiates and absorbs electromagnetic energy, or that such energy carries momentum. Thus it is rather amazing that Kirchhoff should have established on the basis of relatively simple arguments that within a cavity at equilibrium, this radiation should be independent of the substance of the walls of the cavity, and that at a fixed temperature a good emitter of radiation should be a good absorber. A perfect absorber should then radiate an energy equivalent to everything that falls upon it within the cavity at equilibrium, independently at each frequency. He called the radiation emitted by such a perfect absorber the black-body radiation, and postulated that there should be a universal function $u(\nu,\mathrm{T})$ that describes the radiation density in equilibrium with the walls, that on average gets both absorbed and reemitted, at any particular frequency ν and temperature T. The challenge was to find the explicit form of the function $u(\nu,T)$. This search would eventually lead to the birth of quantum mechanics in early twentieth century.

In 1888 Hertz showed the reality of Maxwell waves. In 1893 Wien applied the laws of thermodynamics and electromagnetism to the problem of black-body radiation and succeeded in reducing Kirchhoff's universal function to a function of one variable. That is as far as one can go in classical physics. Wien tackled the problem of including the frequency in the black-body law by considering an adiabatic motion of a wall of the cavity. This induced a Doppler shift on the radiation, while at the same time the wall did work on the radiation.

The Rayleigh–Jeans formula gave results in agreement with the experimental observations at low frequencies; however, it failed miserably at high frequencies. The radiancy, according to Rayleigh and Jeans, is inversely proportional to the fourth power of the frequency, which indicates that at high frequencies the radiancy will approach infinity, thus leading to unphysical results in the ultraviolet region of the spectrum. By 1900, this failure, known as the Rayleigh–Jeans ultraviolet catastrophe, had caused people to question the basic concepts of classical physics and thermodynamics.

It was, however, Max Planck (1858–1947) who would eventually present the radiation formula that matched the experimentally observed black-body radiation spectra for the entire range of frequency spectrum. Planck presented his results that would eventually revolutionize our understanding of the laws of nature, literally at the close of the nineteenth century, on December 15, 1900, at a meeting of the German Physical Society.

When Planck addressed the problem of black-body radiation, he realized that since the results were independent of the nature of the material in the cavity, one could use a simple model for the cavity. So he chose to consider a damped harmonic oscillator as a model for the material in the walls. Planck's derivation consisted of three steps. In the electromagnetic step, he calculated the equilibrium energy of these harmonic oscillators of frequency ν driven by the periodic electric field of frequency ω. In the thermodynamic step, he calculated the entropy of the linear oscillators that gave the correct value for the function $u(\nu,\mathrm{T})$. In the third

1

and crucial statistical step, he calculated the entropy of the linear oscillators and showed that the expression for entropy in the thermodynamic step could only be recovered if he assumed the total energy of the oscillators was made up of finite energy elements, and each element had an energy ε that is equal to $n\hbar\nu$. Here n is an integer and $\hbar$ is a constant that eventually carried Planck's name and is called Planck's constant. This last step was a departure from a classical description and Planck would later describe it as "an act of desperation" to get the correct expression for the Kirchhoff function that agreed with experiments. It is important to realize that the Planck relation $\epsilon = n\hbar\nu$, for integer values of n, is a significant departure from classical thought in two ways. First, it postulates that energy is proportional to frequency, not amplitude, as would be expected for a classical oscillator. Second, for a given frequency ν, the energy is quantized, i.e., it comes in units of $\hbar\nu$.

Planck's derivation for the black-body spectrum was based on the quantization of the material of the cavity and not the radiation itself. However it would have far reaching consequences for the ultimate description of the nature of light through the work of Albert Einstein and others.

Einstein and the Notion of Photon The revival of the particle theory of light, and the beginning of the modern concept of the photon, is due to Albert Einstein (1879–1955). Einstein is a giant in the history of science. He is the founding figure of both quantum mechanics and the theory of relativity. His impact on our understanding of the nature of light is immense.

In his 1905 paper on the photoelectric effect, the emission of electrons from a metallic surface irradiated by UV rays, Einstein was forced to postulate that light comes in discrete bundles, or quanta of energy, borrowing Planck's hypothesis: $\epsilon = n\hbar\nu$. This re-introduced the particulate nature of light into physical discourse, not as localization in space in the manner of Newton's corpuscles, but as discreteness in energy. This gave the Planck hypothesis a new and bold meaning.

There were three issues associated with the photoelectric effect: When light of frequency ν falls on a photoemissive surface, energy of the ejected electron T_e obeys $\hbar\nu = \Phi + T_e$; rate of emission is proportional to the intensity of incident light; and there is no time delay between the time in which the field begins falling on the photoactive surface and the instance of photoelectron emission. The first two of these phenomena can, in contrast to what we read in most textbooks, be explained fully by simply quantizing the atoms associated with the photodetector. However, the third point, namely the lack of a delay is a bit more subtle. It may be reasonably argued that quantum mechanics teaches us that the rate of ejection is finite even for small times, i.e., times involving a few optical cycles of the radiation field. Nevertheless, one may argue that the concept of the photon is really explicit here in the sense that conservation of energy is at stake. That is, if we have only a short period of time τ elapsing between the instants that the radiation field E begins to interact with the photoemitting atoms and the emission of the photoelectron, the amount of energy which has fallen on the surface would be given by $\epsilon_0 E^2 A\tau$, where A is the cross-section of the incident beam. For sufficiently short times, the energy which has fallen on the photodetector may not exceed $\hbar\nu$. This clearly shows that we are not able to conserve energy if we take a semiclassical point of view. However, the photon concept in which the ejection of the photoelectron implies that a photon is annihilated gets around this problem completely. This is one of the triumphs of the quantum field theory. In any case, it is a tribute to Einstein's deep understanding of physics that he was able to introduce the photon concept from such limited, and in some ways, misleading information.

This was a difficult situation. On the one hand, were the interference and diffraction experiments that required a wave nature of light for their explanation

and, on the other, was the photoelectric effect that could be understood by invoking a particle type picture. A complete resolution and the formal theory that would rigorously explain all these phenomena would have to wait almost a quarter century—till the birth of quantum mechanics in the summer of 1925.

Before discussing these developments, we briefly discuss the other "cloud" at the end of the nineteenth century—the null result of the Michelson–Morley experiment—and the birth of the theory of relativity.

Michelson–Morley Experiment and the Birth of the Theory of Relativity Towards the end of nineteenth century, the concept of ether was firmly ingrained within the physics community. For example, Maxwell stated in an article entitled *Ether* for the *Encyclopedia Britannica* (1878): "There can be no doubt that the interplanetary and intersteller spaces are not empty but are occupied by a material substance or body, which is certainly the largest, and probably the most uniform, body of which we have any knowledge." He himself attempted unsuccessfully to measure the influence of ether's drag on the motion of the earth. It was, however, Albert Michelson (1852–1931) and Edward Morley (1838–1923), who carried out an experiment in 1887 to decisively establish the existence of Ether. They sent white light, through a half-silvered mirror into an interferometer, now called Michelson interferometer. The light beam was split into two beams, one of them traveling straight to a mirror in one arm and the other propagating at right angles to another mirror, with both beams recombining at the beam splitter after traveling equal distances They thus produced a pattern of constructive and destructive interference whose transverse displacement would depend on the relative time the light took to traverse the paths in the two arms. If the earth moves through the ether medium, the beam traveling along ether would take a longer time than the beam traveling in the perpendicular direction. Michelson and Morley expected a fringe shift equal to 0.4 fringes. What they measured was the maximum displacement of 0.02 and an average shift much less than 0.01. They thus concluded that the hypothesis concerning the existence of Ether medium is false. This null result—the most famous null result in the history of physics—was initially a major disappointment.

A resolution of the null result of the Michelson–Morley experiment came in 1889 by an Irish physicist, George FitzGerald. He postulated that the results of Michelson–Morley experiment could be explained using the hypothesis of the contraction of moving bodies in the direction of motion, the amount of contraction being just the right amount to give the same time difference as to explain the null result. According to him: "I would suggest that almost the only hypothesis that can reconcile. . . is that the length of the material bodies changes, according as they are moving through the ether or across it, by an amount depending on the square of the ratio of their velocities to that of light." In 1892, unaware of FitzGerald's hypothesis, Lorentz came to the same conclusion. Today we call it FitzGerald–Lorentz contraction. This was, however, an ad-hoc solution to the null result of the Michelson–Morley experiment with no basis in theory. This set up Lorentz (1853–1928) on the road to the derivation of Lorentz transformation and Einstein to his theory of relativity.

In 1905, Einstein formulated the theory of special relativity, based on the two postulates: The principle of relativity, i.e., the laws of physics do not change, even for objects moving in inertial (constant speed) frames of reference and the principle of the speed of light, i.e., the speed of light c is the same for all observers, regardless of their motion relative to the light source. Based on these postulates, Einstein could derive the Lorentz transformation and the length contraction. This represented a major departure from the Newton's notion of absolute space and

1

time. A most celebrated consequence of the theory of relativity was the equivalence of energy E and mass m via the relation $E = mc^2$. The notion of stationary ether that had been the ever existing background in all the theories since antiquity played no role in the theory of relativity.

Einstein also went on to develop a general theory of relativity that would provide a geometric theory of gravitation. This theory is based on the equivalence principle under which the states of accelerated motion and being at rest in a gravitational field are physically identical. In 1915, Einstein published a paper in which he described gravity as a geometric property of space and time. In particular, the curvature of spacetime changes in the vicinity of a massive object. The predictions of this theory were at variance with Newton's theory of gravitation and motivated one of the most dramatic experiments in the history of Physics—an experiment that would pit the two giants of science, Isaac Newton and Albert Einstein, and their conflicting theories of gravitation against each other. The experiment was the bending of light by a massive object.

Bending of Light, Newton, Einstein, and Eddington We recall that Newton championed a corpuscular nature for light. Newton had also noted, while formulating his theory of gravitation, that any material particle moving at a finite speed would experience a force while passing in the vicinity of a massive object. This pull by gravity should bend the trajectory of the particle and the bending angle should be independent of the mass of the particle. Thus if light is composed of small particles, they should also experience such deflection. Newton himself did not calculate this deflection as, in his time, the finite speed of light was not well established. However he postulated this deflection and, towards the end of his treatise *Opticks* (1704), he noted "Do not Bodies act upon Light at a distance, and by their action bend its Rays, and is not this action strongest at the least distance?". The finite speed of light was well established by early nineteenth century and a German astronomer, Johann Georg von Soldner (1804), presented calculations based on Newton's corpuscular theory that light weighs and bends like high speed projectiles in a gravitational field. He produced a value of 0.87 arc sec bending angle for light grazing the Sun.

In 1911, more than hundred years later, Einstein calculated the bending of light by combining the equivalence principle with special theory of relativity to predict a deflection of light from the sun by the angle of 0.87 arc second. This is the same value that Newtonian theory predicted. He obtained this result before he formulated the general theory of relativity and the associated curved space time. When he included the effects of general theory of relativity, the predicted value for the bending of light doubled to 1.83 arc second. This result was published on November 18, 1915. Thus the predictions of Newton and Einstein were at odds with each other and an experimental activity followed soon to decide who was right. The bending of light by a massive object also became the first test of the esoteric Einstein's general theory of relativity.

After the First World War was over, Sir Arthur Eddington (1882–1944) organized an expedition to the island of Principe near Africa to watch the solar eclipse on May 29, 1919 and to measure the observed curving of light from distant stars by the gravitational pull of the sun. While the expedition was being planned, Eddington wrote: "The present eclipse expeditions may for the first time demonstrate the weight of light (i.e. Newton's value) and they may also confirm the added effect of Einstein's weird theory of non-Euclidean space, or they may lead to a result of yet more far reaching consequences of no deflection". When the results were announced, they agreed with Einstein's predicted value. Einstein became an overnight international celebrity and an iconic figure.

Birth of Quantum Mechanics The first quarter of twentieth century was perhaps the most remarkable period in the history of Physics. Through the discoveries of the quantum theory of Planck, Einstein, and Niels Bohr and the Einstein's theories of special and general relativity, the outlook on conventional or classical Physics had completely transformed. Newtonian Physics was unable to explain effects that happened at sub-atomic level or at high speeds, speeds comparable to the speed of light. The capstone of these developments was the birth of quantum mechanics that took place in the summer and winter of 1925 through the works of Werner Heisenberg, Max Born, Pascual Jordan, and Paul Dirac, on one hand, and Erwin Schrodinger, on the other. This new theory, that replaced Newton's and Maxwell's theories, would have revolutionary consequences in our story on the nature of light. An important underlying feature of the new theory was the notion of complementarity, namely two observables are *complementary* if precise knowledge of one of them implies that all possible outcomes of measuring the other one are equally probable. This injected the notion of wave-particle duality in the discourse on the nature of both light and matter.

Dirac, Quantum Theory of Light With the advent of quantum mechanics, the dual nature of light was apparent. There were phenomena such as interference and diffraction that could be explained based on the wave nature of light. Then there were phenomena such as excitation of an atom by absorbing a photon that required a particle nature of light. It was Paul Adrien Dirac (1902–1982) who, in a seminal paper published in 1927, synthesized the wave and particle natures of light in a single theory. According to the Maxwell's theory, the light consisted of electromagnetic waves of different frequencies. The oscillating waves could be looked upon as a sort of simple harmonic oscillators. Central to Dirac's quantum theory of radiation was the notion that each mode of the electromagnetic field could be identified as a quantized simple harmonic oscillator. Both satisfy the same commutation relation $[\widehat{q}, \widehat{p}] = i\hbar$, although q and p represent different things in the two cases. In the case of harmonic oscillator, they represent the position and momentum of the oscillating particle, while in the case of electromagnetic case, they represent the electric (E) and magnetic (B) fields of the light in a given wavevector and polarization mode k. Thus, the quantum electromagnetic field consists of an *infinite* product of such generalized harmonic oscillators, one for each mode of the field. A Heisenberg-type uncertainty relation applies to the Maxwell fields: $\Delta E \Delta B \geq \hbar/2 \times$ constant, i.e., the electric and the magnetic fields associated with light cannot be measured arbitrarily precisely. Such field fluctuations are an intrinsic feature of the quantized theory. The uncertainty relation can also be formulated in terms of the in-phase and in-quadrature components of the electric field. To introduce the notion of a photon, it is convenient to recast the above quantization of the field in terms of the annihilation ($\hat{a}$) and creation ($\hat{a}^\dagger$) operators of a harmonic oscillator. These correspond to the positive and negative frequency parts of the electric field operator, respectively.

By analogy to the theory of the harmonic oscillator, the application of $\hat{a}$ produces a state with one *less* quantum of energy, and the application of $\hat{a}^\dagger$ produces a state with one *more* quantum of energy. This naturally leads to discrete energies for the oscillator in each mode: $n_k = 0, 1, 2, \ldots$. In the absence of any medium, the modes $E_k(\mathbf{r})$ are just the plane-wave solutions to the Maxwell equations. Alternately, we can define a localized "pulse" basis for the photon by summing over many wave vectors and frequencies, just as for classical waves. Thus, quantum electrodynamics permits both wave and particle perspectives on

1

light. The wave perspective is exemplified by the picture of a stochastic electromagnetic field. The particle perspective follows from the language of annihilation and creation operators which are subject to the appropriate commutation relation. Combining these perspectives, one can adopt a rigorous definition of the photon as follows: A photon corresponds to a single excitation of a particular mode k of the electromagnetic field in a suitably defined cavity, such that the annihilation and the creation operators for the field mode satisfy a Boson commutation relation. The wave-particle picture of light embedded in the Dirac's theory of light had novel and important consequences. The most important was the reshaping of our concept of vacuum.

Quantum Vacuum Before the advent of quantum field theory, the vacuum was perceived as *nothing*—a place where no light existed, nothing moved, and there was no energy present. The quantum mechanical picture of vacuum turned out to be dramatically different. According to Dirac's theory of light, the quantum harmonic oscillator associated with each electromagnetic wave of frequency ν has an energy equal to $\hbar\nu/2$ in vacuum. There are infinite number of mode in the universe, each associated with a frequency ν. Thus the total energy in the universe can be calculated by adding this vacuum energy for each mode and the result is an infinite amount of energy. In addition, as noted above, there are quantum mechanical fluctuations as a result of Heisenberg's uncertainty relation that cannot be neglected, even in vacuum. This forbids a classical description of absolutely zero electric and magnetic fields in vacuum. Instead we have fluctuations—randomly nonzero fields at any time. We thus have a revolutionary new way of thinking about light that field quantization introduced into the scientific discourse, namely that the electromagnetic field, when quantized, has the ability to exist in a state of pure nothingness—the so-called vacuum state—and yet have observable consequences in the material world.

Spontaneous Emission An important consequence of the fluctuations in the vacuum field is the phenomenon of spontaneous emission by an atom. A photon is created in response to these fluctuations. Thus, even in the absence of an applied field, an atom in the excited state can decay to the ground state and spontaneously emit a photon. Since the direction and time of emission are random, this process represents a fundamental source of quantum noise, and a limitation to any coherent process (such as lasing). The excited atomic level acquires a finite bandwidth which is the inverse of the emission lifetime. We can use quantum theory to calculate the spatio-temporal profile of the emitted photon as detected by a photodetector.

Lamb Shift Perhaps the greatest triumph of field quantization is the explanation of the Lamb shift between, for example, the $2s_{1/2}$ and $2p_{1/2}$ levels in a hydrogenic atom. Relativistic quantum mechanics predicts that these levels should be at the same energy. Willis Lamb (1913–2008), however, experimentally observed in 1947 a frequency splitting of about 1 GHz in contradiction to the theoretical prediction. We can understand the shift intuitively by picturing the electron forced to fluctuate about its first-quantized position in the atom due to random kicks from the surrounding, fluctuating vacuum field. Its average displacement $< \Delta\mathbf{r} >$ is zero, but the squared displacement $< (\Delta\mathbf{r})^2 >$ is slightly nonzero, with the result that the electron "senses" a slightly different Coulomb pull from the positively charged nucleus than it normally would. The effect is more prominent nearer the nucleus where the Coulomb potential falls off

more steeply, thus the s orbital is affected more than the p orbital. This is manifested as the Lamb shift between the levels.

Casimir Force In 1947, Hendrick Casimir (1909–2000) predicted that if two conducting plates separated by a distance a are placed in vacuum, and no external force is acting on them, they would attract each other with a force equal to $\hbar c\pi^2/240a^4$. This Casimir force was experimentally observed in 1958. Casimir explained this force arising purely as a consequence of the quantized modes of the radiation in vacuum. When the two conducting plates are inserted in the vacuum, the space is divided into three regions, the two infinite regions outside the plates and another region inside the two plates. The regions outside the plates have continuum of frequencies, i.e., all possible frequencies, resulting in an infinite amount of energy when we add the contributions of all the modes. The region inside the plates, however, allows only discrete number of modes satisfied by the resonance condition $a = \pi nc/\nu_n$, where ν_n is the frequency of the nth mode. These are also infinite number of modes, one for each value of n. The total amount of the vacuum field energy between the plates is also infinite. Thus we have an infinite amount of energy outside the plates and an infinite amount of energy between the plates. The truly dramatic result is that when we subtract these two infinities, the outcome is finite. As the system tends to evolve to a state with minimum energy there is a resulting force and this force is attractive. This is a highly counterintuitive result. Julian Schwinger called it "One of the least intuitive consequences of quantum electrodynamics" and according to Bryce DeWitt: "What startled me, in addition to the crazy idea that a pair of electrically neutral conductors should attract one another, was the way in which Casimir said the force could be computed, namely, by examining the effect on the zero-point energy of the electromagnetic vacuum caused by the mere presence of the plates. I had always been taught that the zero-point energy of a quantized field was unphysical."

Laser: A Coherent Light Source All the studies on light from the antiquity until the middle of nineteenth century were based on incoherent light sources such as the sun, candle light, sodium lamp, or light bulb. In 1950s a new coherent source of light was invented, first in the microwave region and then in the optical region. This new kind of light source, laser, is one of the greatest inventions of the second part of the twentieth century. It has helped to revolutionize many branches of science and technology ranging from bio-technology and precision measurements to communication and remote sensing. The physical process behind conventional light sources is spontaneous emission and they operate in thermal equilibrium. Initially majority of atoms and molecules are in their ground state. When energy is supplied to the atoms or molecules, some of them go to the excited states and then radiate via spontaneous emission. As discussed above, the spontaneous emission process is due to the ubiquitous vacuum fluctuations and each atom radiates independently of each other. The resulting light is a white light sent in all directions and is incoherent. On the other hand, the dominant emission process in a laser is stimulated emission. By a clever design, the radiated photons by the atoms or molecules are able to stimulate other atoms to radiate with the same frequency and same direction. The resulting radiation is coherent, monochromatic, and highly directional.

In 1954, Gordon, Zeiger, and Charles Townes (1915–2015) showed that coherent electromagnetic radiation can be generated in the radio frequency range by the so-called maser (microwave amplification by stimulated emission of

radiation). The first maser action was observed in ammonia. The maser principle was extended by Arthur Schawlow (1921–1999) and Townes, and also by Basov and Prokhorov, to the optical domain, thus obtaining a LASER (light amplification by stimulated emission of radiation). A laser consists of a set of atoms interacting with an electromagnetic field inside a cavity. The cavity supports only a specific set of modes corresponding to a discrete sequence of frequencies. The active atoms, i.e., the ones that are pumped to the upper level of the laser transition, are in resonance with one of these frequencies of the cavity. A resonant electromagnetic field gives rise to stimulated emission, and the atoms transfer their excitation energy to the radiation field. The emitted radiation is still at resonance. If the upper level is sufficiently populated, this radiation gives rise to further transitions in other atoms. In this way all the excitation energy of the atoms is transferred to a single mode of the radiation field.

The first pulsed laser operation was demonstrated by Theodore Maiman (1927–2007) in ruby in 1960. The first continuous wave (cw) laser, a He–Ne gas laser, was built by Ali Javan later in the same year. Since then, a large variety of systems have been demonstrated to exhibit lasing action; generating coherent light over a frequency domain ranging from infrared to ultraviolet. These include dye lasers, chemical lasers, and semiconductor lasers.

The Birth of Quantum Optics The advent of laser required a careful description of the various sources of light. The question was: What is the fundamental difference between the conventional light sources, such as the sun, and the newly discovered laser light? An answer to this question led to a new field of study in Physics: Quantum Optics. Roy Glauber, in a series of seminal papers in 1963, at first controversial, differentiated between laser (coherent) light and normal (blackbody) light in terms of the photon statistics. This work had far reaching consequences as it showed that there could be all kinds of light sources that have to be distinguished by their quantum states and the corresponding statistical properties. These sources could range from a photon number state, in which light quanta could behave like particles, to a coherent state, which is as close to a classical Maxwellian description of light as electromagnetic wave as the quantum mechanics allows.

Quantum Interference and Delayed Choice Quantum Eraser Before closing this brief story of light, we observe that the paradigm of quantum interference, the interference of probability amplitudes associated with different paths taken by a photon, defines our present understanding on the nature of light. In some ways, this is a culmination of the centuries-old debate on the nature of light reviewed through this article. The modern quantum perspective on this debate is that light is neither wave nor particle, but an elusive, intermediate entity that obeys the superposition principle. The quintessential experiment that demonstrates wave-particle duality is the Young's two-slit interference experiment. When a single photon goes through the slits, it registers as a point-like event on the screen (measured say by a CCD array). An accumulation of such events over repeated trials builds up a probabilistic fringe pattern that is characteristic of wave interference. However, if we arrange to measure which slit the photon goes through, the interference always disappears.

This picture is, however, not so simple. The counterintuitive aspect of complementarity is epitomized in the problem of quantum eraser, as was shown by Marlan Scully in 1982. The inability to discern which-path information, or the indistinguishability of interfering pathways, in the double-slit experiment is the key to preserving the wave properties of the photon and the appearance of fringes

on the screen. What if, rather than subject the photon to a classical measurement, we can have it interact *quantum mechanically* with a localized marker particle (such as an atom) and leave behind a record of its path? The interference pattern then survives or not depends on the marker states, which carry the tell-tale information about which path the photon took to the detector. The coherence is destroyed as soon as we have the which-path information. One then wonders whether it might not be possible to retrieve the coherence, and the fringes, by destroying the which-path information contained in the marker—long after the photon is detected on the screen. This is the essence of the quantum eraser idea. An experimental demonstration of quantum eraser elicits a response of incredulity as the following quote by Brian Greene in his beautiful book *The Fabric of the Cosmos* indicates: "These experiments are a magnificent affront to our conventional notions of space and time. . . . . For a few days after I learned of these experiments, I remember feeling elated. I felt I'd been given a glimpse into a veiled side of reality."

1.6 Epilogue

We have come a long way from the earliest studies on light, trying to understand vision as light emanating from our eyes, to the description of light as rays, then as particles, and then waves, and finally exhibiting both particle and wave natures. We can only speculate how our present understanding of light will be perceived decades or centuries from now. Will our picture of light quanta as both waves and particles survive or will something more intuitive replace this incomprehensible picture? It is an irony that the greatest strides taken in the scientific understanding have come in our time, yet we feel least certain of our understanding of what light is, what photon is. In spite of the great success of the mathematical theory to describe light and its amazing agreement with experiment, the question "What is Light?" can ignite a heated discussion. To quote Albert Einstein (1954), "All the fifty years of conscious brooding have brought me no closer to the answer to the question: What are light quanta? Of course today every rascal thinks he knows the answer, but he is deluding himself."

Bibliography

Lindberg DC (1976) Theories of vision: from Al-Kindi to Kepler. University of Chicago Press, Chicago

Pais A (1982) Subtle is the lord: the science and the life of Albert Einstein. Oxford University Press, New York

Scully MO, Suhail Zubairy M (1997) Quantum optics. Cambridge University Press, Cambridge
Aharonov Y, Suhail Zubairy M (2005) Time and the quantum: erasing the past and impacting the future. Science 307:875
Muthukrishnan A, Scully MO, Suhail Zubairy M (2003) The concept of the photon-revisited? Opt Photonics News 14(10):18–27
Greenberger DM, Erez N, Scully MO, Svidzinsky AA, Suhail Zubairy M (2007) Planck, photon statistics, and Bose-Einstein condensate. In: Wolf E (ed) Progress in optics, vol 50. Elsevier, Amsterdam, p 275
Lyons J (2009) The house of wisdom: how the Arabs transformed western civilization. Bloomsbury Press, New York
Darrigol O (2012) A history of light: from Greek antiquity to the nineteenth century. Oxford University Press, Oxford
Mark Smith A (2015) From sight to light: the passage from ancient to modern optics. University of Chicago Press, Chicago
Weinberg S (2015) To explain the world: the discovery of modern science. HarperCollins, New York

Ibn al-Haytham's Scientific Research Programme

Roshdi Rashed

R. Rashed (✉)
Université Paris Diderot, Sorbonne Paris Cité, SPHERE, UMR 7219, CNRS, 5 rue Thomas Mann, Bâtiment Condorcet, Case 7093, 75205 Paris Cedex 13, France
e-mail: rashed@paris7.jussieu.fr

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_2

2

2.1 Introduction

For the vast majority of historians, and, more generally, of laymen, Ibn al-Haytham's major contribution concerns the vision in all its aspects (physical, physiological and psychological) and, namely, the causes of perceptual and cognitive effects. The reform of Ibn al-Haytham, according to them, was mainly to abandon the traditional theory of vision, to a new one. Henceforth he belongs to ancient and mediaeval traditions, in spite of this reform, in so far that he was concerned with vision and sight.

I will argue here that this reform was a minor consequence of a more general and more fundamental research programme, and even his conception of the science of optics is quite different as so far that his main task was about light, its fundamental properties and how they determine its physical behaviour, as reflection, refraction, focalization, etc.

Some historians of optics consider that, up to the seventeenth century in Europe, the science in optics before Kepler was aimed primarily at explaining vision. The merest glance at the optical works of Ibn al-Haytham leaves no doubt that this global judgement is far from being correct. Indeed, this statement is correct as far as it concerns the history of optics before the shift done by Ibn al-Haytham and the reform he accomplished. Successor of Ptolemy, al-Kindī and Ibn Sahl, to mention only a few, he unified the different branches of optics: optics, dioptrics, anaclastics, meteorological optics, etc. This unification was possible only for a mathematician who focused on light, and not on vision. Nobody, as far as I know, before Ibn al-Haytham, wrote such books titled: *On Light*; *On the Light of the Moon*; *On the Light of the Stars*; *On the Shadows*, among others, in which nothing concerns sight. At the same time, three books from his famous *Book of Optics* are devoted strictly to the theory of light. None of the authors before him, who were mainly interested in vision, wrote a very important contribution on physical optics such as the one on *The Burning Sphere*.

I begin by quoting the expression which Ibn al-Haytham repeated more than once in his different writings on optics. At the beginning of this famous *Book of Optics*, he writes:

> » Our subject is obscure and the way leading to knowledge of its nature difficult, moreover our inquiry requires a combination of the natural and mathematical sciences.[1]

But such a combination in optics, for instance, requires one to examine the entire foundations and to invent the means and the procedures to apply mathematics on the ideas of natural phenomena. For Ibn al-Haytham, it was the only way to obtain a rigorous body of knowledge.

Why this particular turn, at that time? Let me remind that Ibn al-Haytham lived in the turn of the first millennium. He was the heir of two centuries of scientific research and scientific translations, in mathematics, in astronomy, in statics, in optics, etc. His time was of intense research in all these fields. He himself wrote more in mathematics and in astronomy than in optics per se. According to early bio-bibliographers, Ibn al-Haytham wrote 25 astronomical works: twice as many works on the subject as he did in optics. The number of his writings alone indicates the huge size of the task accomplished by him and the importance of astronomy in his life work. In all branches of mathematics, he wrote more than all his writings in astronomy and in optics put together. If he wrote in optics the famous huge book, *Kitāb al-Manāẓir—The Book of Optics*, in astronomy likewise

1 Ibn al-Haytham [2], p. 4.

he wrote a huge book entitled *The Configuration of the Motions of each of the Seven Wandering Stars*.

Before coming back in some details to these contributions, let me characterize Ibn al-Haytham's research programme.

1. It is a new one, concerning the relationships between mathematics and natural phenomena, never conceived before. His aim is to mathematize every empirical science. This application of mathematics can take different forms, not only given to the different disciplines, but also in one and the same discipline.
2. It does not concern only optics, but every natural science, i.e., for the epoch, astronomy and statics.
3. Its success depends on the means—mathematical, linguistic and technical—by which mathematics control the semantic and syntactical structures of natural phenomena.

2.2 Between Ptolemy and Kepler: Ibn al-Haytham's Celestial Kinematics

To put the facts right, I will turn at first, quite briefly, to Ibn al-Haytham's astronomy. He wrote at least three books criticizing the astronomical theory of Ptolemy:

1. *The Doubts concerning Ptolemy*
2. *Corrections to the* Almagest
3. *The Resolution of Doubts concerning the* Almagest

In the *Doubts*, Ibn al-Haytham comes to the conclusion that "the configuration Ptolemy assumes for the motions of the five planets is a false one".[2] A few lines further on, he continues: "The order in which Ptolemy had placed the motions of the five planets conflicts with the theory <that he had proposed>".[3] A little later, he states: "The configurations that Ptolemy assumed for the <motions of> the five planets are false ones. He decided on them knowing they were false, because he was unable <to propose> other ones.[4]" After such comments, and many others like them in several places of his writings, Ibn al-Haytham had no option but to construct a planetary theory of his own, on a solid mathematical basis, and free from the internal contradictions found in Ptolemy's *Almagest*. For this purpose, he conceived the idea of writing his monumental and fundamental book *The Configuration of the Motions of the Seven Wandering Stars*. If we wish to characterize the irreducible inconsistencies that, according to Ibn al-Haytham, vitiate Ptolemy's astronomy, we may say that they arise from the poor fit between a mathematical theory of the planets and a cosmology; that is, the combination between mathematics and physics. Ibn al-Haytham was familiar with similar, though of course not identical, situations when, in optics, as we shall see, he encountered the inconsistency between geometrical optics and physical optics as understood not only by Euclid and Ptolemy, but also by Aristotle and the philosophers.

In *The Configuration of the Motions* he deals with the apparent motions of the planets, without ever raising the question of the physical explanation of these motions in terms of dynamics. It is not the causes of celestial motions that interest Ibn al-Haytham, but only the motions themselves observed in space and time. Thus, to proceed with the systematic mathematical treatment, and to avoid the

2 See Rashed [8], p. 13.

3 See footnote 2.

4 See footnote 2.

2

obstacles that Ptolemy had encountered, he first needed to break away from any kind of cosmology. Thus the purpose of Ibn al-Haytham's *Configuration of the Motions* is clear: instead of constructing, as his predecessors, a cosmology, or a kind of dynamics, he constructs the first geometrical kinematics.

A close examination of the way he organizes his exposition of planetary theory shows that Ibn al-Haytham begins by omitting physical spheres and by proposing simple—in effect, descriptive—models of the motions of each of the seven planets. As the exposition progresses, he makes the models more complicated and increasingly subordinates them to the discipline of mathematics. This growing mathematization leads him to regroup the motions of several planets under a single model. This step obviously has the effect of privileging a property that is common to several motions. In this way Ibn al-Haytham opens up the way to achieving his principal objective: to establish a system of celestial kinematics. He does so without as yet formulating the concept of instantaneous speed, but by using the concept of mean speed, represented by a ratio of arcs.

In the course of his research, which I analysed elsewhere,[5] we encounter a concept of astronomy that is new in several respects. Ibn al-Haytham sets himself the task of describing the motions of the planets exactly in accordance with the paths they draw on the celestial sphere. He is neither trying 'to save the phenomena', like Ptolemy, that is, to explain the irregularities in the assumed motion by means of artifices such as the equant; nor trying to account for the observed motions by appealing to underlying mechanisms or hidden natures. He wants to give a rigorously exact description of the observed motions in terms of mathematics. Thus his theory for the motion of the planets calls upon no more than observation and conceptual constructs susceptible of explaining the data, such as the eccentric circle and in some cases the epicycle. However, this theory does not aim to describe anything beyond observation and these concepts, and in no way is it concerned to propose a causal explanation of the motions.

The new astronomy no longer aims at constructing a model of the universe, as in the *Almagest*, but only at describing the apparent motion of each planet, a motion composed of elementary motions, and, for the inferior planets, also of an epicycle. Ibn al-Haytham considers various properties of this apparent motion: localization and the kinematic properties of the variations in speed.

In this new astronomy, as in the old one, every observed motion is circular and uniform, or composed of circular and uniform motions. To find these motions, Ibn al-Haytham uses various systems of spherical coordinates: equatorial coordinates (the required time and its proper inclination); horizon coordinates (altitude and azimuth) and ecliptic coordinates. The use of equatorial coordinates as a primary system of reference marks a break with Hellenistic astronomy. In the latter, the motion of the orbs was measured against the ecliptic, and all coordinates were ecliptic ones (latitude and longitude). Thus, basing the analysis of the planets' motion on their apparent motions drives a change in the reference system for the data; we are now dealing with right ascension and declination. Ibn al-Haytham's book thus transports us into a different system of analysis.

To sum up, in the *The Configuration of the Motions*, Ibn al-Haytham's purpose is purely kinematics; more precisely, he wanted to lay the foundations of a completely geometrical kinematics tradition. But carrying out such a project involves first of all developing some branches of geometry, as also of plane and spherical trigonometry. In both fields, Ibn al-Haytham obtained new and important results.

In astronomy, properly, there are two major processes that are jointly involved in carrying through this project: freeing celestial kinematics from cosmological

5 See Rashed [8].

connections, that is, from all considerations of dynamics, in the ancient sense of the term; and to reduce physical entities to geometrical ones. The centres of the motions are geometrical points without physical significance; the centres to which speeds are referred are also geometrical points without physical significance; even more radically, all that remains of physical time is the 'required time', that is, a geometrical magnitude. In short, in this new kinematics, we are concerned with nothing that identifies celestial bodies as physical bodies. All in all, though it is not yet that of Kepler, this new kinematics is no longer that of Ptolemy nor of any of Ibn al-Haytham's predecessors; it is *sui generis*, half way between Ptolemy and Kepler. It shares two important ideas with ancient kinematics: every celestial motion is composed of elementary uniform circular motions, and the centre of observation is the same as the centre of the Universe. On the other hand, it has in common with modern kinematics the fact that the physical centres of motions and speeds are replaced by geometrical centres.

In fact, once Ibn al-Haytham had engaged upon mathematizing astronomy and had noted not only the internal contradictions in Ptolemy, but doubtless also the difficulty of constructing a self-consistent mathematical theory of material spheres using an Aristotelian physics, he conceived the project of giving a completely geometrized kinematic account.

Ibn al-Haytham had the same experience in optics. In astronomy, kinematics and cosmology are entirely separated to effect a reform of the discipline, just as in optics, work on light and its propagation is entirely separated from work on vision to effect a reform of optics; in the one case as in the other, we shall see, Ibn al-Haytham arrived at a new idea of the science concerned.

2.3 Ibn al-Haytham's Reform of Optics

It is now time to come to Ibn al-Haytham's optics. As we have said above, Ibn al-Haytham was preceded by two centuries of translation into Arabic of the main Greek optical writings, as well of inventive research. Among his Arabic predecessors, al-Kindī, Qusṭā ibn Lūqā, Aḥmad ibn ʿĪsā ʿUṭārid, etc. During these two centuries, the interest shown in the study of burning mirrors is an essential part of the comprehension of the development of catoptrics, anaclastics and dioptrics, as the book produced between 983 and 985 by the mathematician al-ʿAlā' ibn Sahl testifies. Before this contribution of Ibn Sahl, the catoptricians like Diocles, Anthemius of Tralles, al-Kindī etc.[6] asked themselves about geometrical properties of mirrors and about light they reflect at a given distance. Ibn Sahl modifies the question by considering not only mirrors but also burning instruments, i.e. those which are susceptible to light not only by reflection, but also by refraction; and how in each case the focalization of light is obtained. Ibn Sahl studies then, according to the distance of the source (finite or infinite) and the type of lighting (reflection or refraction) the parabolic mirror, the ellipsoidal mirror, the plano-convex lens and the biconvex lens. In each of these, he proceeds to a mathematical study of the curve, and, then, expounds a mechanical continuous drawing of it. For the plano-convex lens, for instance, he starts by studying the hyperbola as a conic section, in order then to take up again a study of the tangent plane to the surface engendered by the rotation of the arc of hyperbola around a fixed straight line, and, finally, the curve as an anaclastic curve, and the laws of refraction.

These studies which focused on light and its physical behaviour were instrumental in the discovery by Ibn Sahl of the concept of a constant ratio, characteristic

6 See Rashed [5, 6].

2

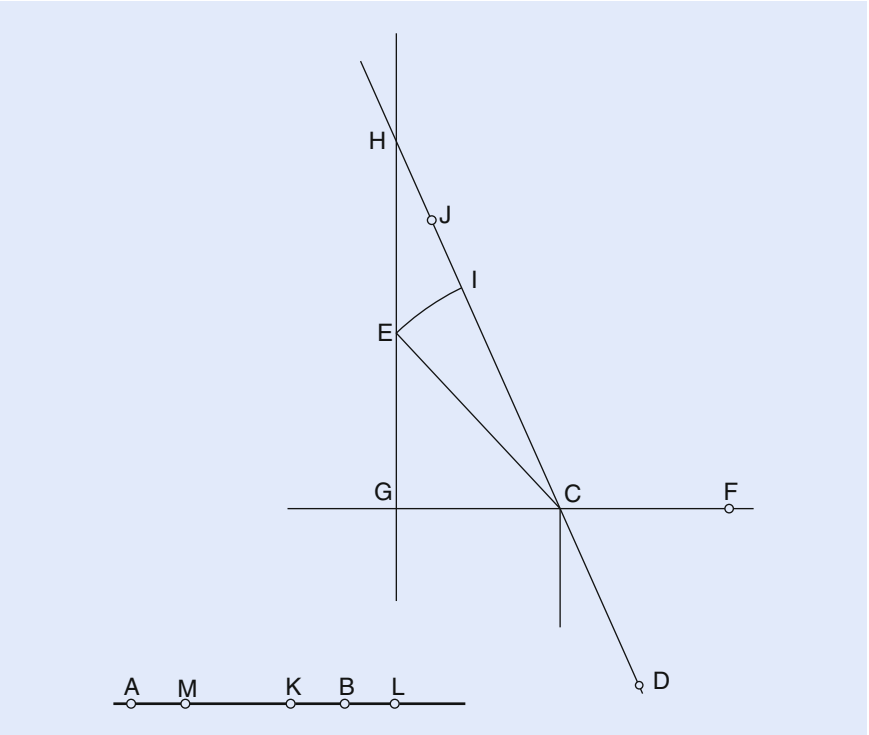


Fig. 2.1 Ibn Sahl illustration of a ray of light (*DC*) refracted as it crosses the boundary (*GF*) of two media of different refractive indices (see text for more details)

of the medium, which is a masterpiece in his study of refraction in lenses, as well as his discovery of the so-called Snellius' law.

At the beginning of his study, Ibn Sahl considers a plane surface *GF* surrounding a piece of transparent and homogeneous crystal. He next considers the straight line *CD* along which the light propagates in the crystal, the straight line *CE* along which it refracts itself in the air, and the normal at *G* on the surface *GF* which intersects the straight line *CD* at *H* and the ray refracted at *E*.

Obviously, Ibn Sahl is here applying the known law of Ptolemy according to which the ray *CD* in the crystal, the ray *CE* in the air and the normal *GE* to the plane surface of the crystal are found in the same plane (Fig. 2.1). He writes then, in a brief way, and, according to his habit, with no conceptual commentary:

» Straight line *CE* is therefore smaller than straightline *CH*. From straight line *CH*, we separate the straight line *CI* equal to straight line *CE*; we divide *HI* into two halves at point *J*; we make the ratio of straight line *AK* to straight line *AB* equal to the ratio of straight line *CI* to straight line *CJ*. We draw the line *BL* on the prolongation of straight line *AB* and we make it equal to straight line *BK*.[7]

In these few phrases, Ibn Sahl draws the conclusion first that $CE/CH < 1$, which he will use throughout his research into lenses made in the *same* crystal. In effect he does not fail to give this same ratio again, nor to reproduce this same figure, each time that he discusses refraction in this crystal (Fig. 2.2).

7 See Rashed [7], p. 106.

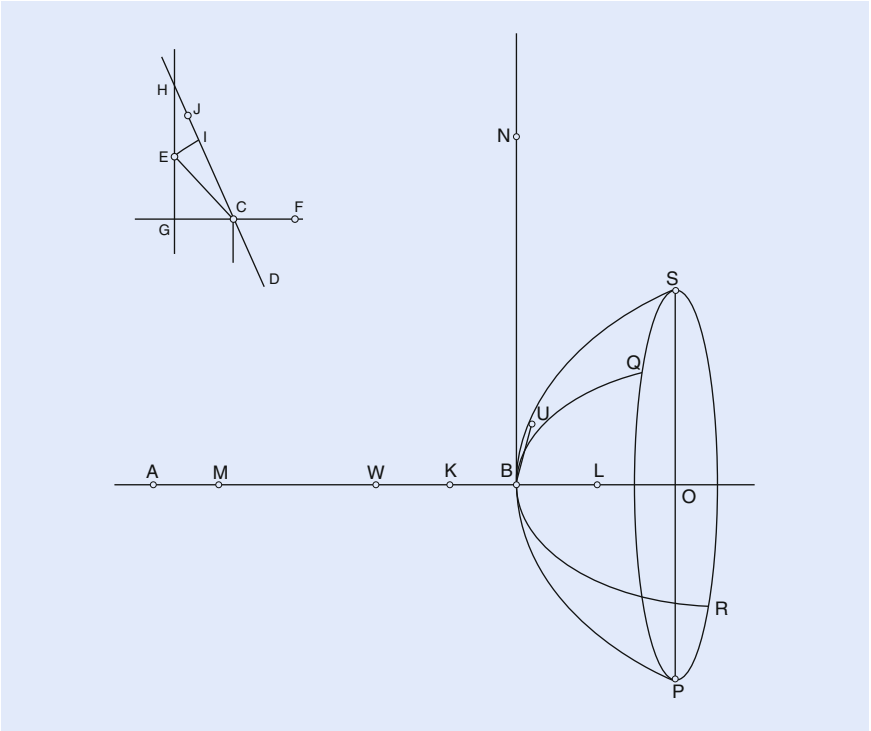


Fig. 2.2 Ibn Sahls diagram depicting refraction with plano-convex lenses (see text for more details)

But the ratio is nothing other than the inverse of the index of refraction in this crystal in relation to the air. Considering the i_1 and i_2 as the angles formed, respectively, by *CD* and by *CE* with the normal *GH*, we have

$$\frac{1}{n} = \frac{\sin i_1}{\sin i_2} = \frac{CG}{CH} \cdot \frac{CE}{CG} = \frac{CE}{CH}.$$

Ibn Sahl takes on the segment *CH* a point *I* such that *CI*=*CE*, and a point *J* at the midpoint of *IH*. This gives

$$\frac{CI}{CH} = \frac{1}{n}.$$

The division *CIJH* characterizes this crystal for all refraction.

Ibn Sahl shows, moreover, in the course of his research into the plano-convex lens and the biconvex lens, that the choice of hyperbola to fashion the lens depends on the nature of the crystal, since the eccentricity of the hyperbola is $e = 1/n$.

Thus, Ibn Sahl had conceived and put together an area of research into burning instruments and, also, anaclastics. But, obliged to think about conical figures other than the parabola and the ellipse—the hyperbola for example—as anaclastic curves, he was quite naturally led to the discovery of the law of Snellius. We understand therefore that dioptrics, when it was developed by Ibn Sahl, only dealt with matters involving the propagation of light, independently of problems of vision. The eye did not have its place within the area of burning instruments, nor did the rest of the subject of vision. It is thus an objective point of view which is deliberately adopted in the analysis of luminous phenomena. Rich in technical material, this new discipline is in fact very poor on physical content: it is evanescent and reduces a few energy considerations. By way of example, at least in his

writings that have reached us, Ibn Sahl never tried to explain why certain rays change direction and are focused when they change medium: it is enough for him to know that a beam of rays parallel to the axis of a plano-convex hyperbolic lens gives by refraction a converging beam. As for the question why the focusing produces a blaze, Ibn Sahl is satisfied with a definition of the luminous ray by its action of setting ablaze by postulating, as did his successors elsewhere for much longer, that the heating is proportional to the number of rays.

Whilst Ibn Sahl was finishing his treatise on *Burning Instruments* very probably in Baghdad, Ibn al-Haytham was probably beginning his scientific career. It would not be surprising therefore if the young mathematician and physicist had been familiar with the works of the elder, if he cited them and was inspired by them. The presence of Ibn Sahl demolishes straightaway the image carved by historians of an isolated Ibn al-Haytham whose predecessors were the Alexandrians and the Byzantines: Euclid, Ptolemy and Anthemius of Tralles. Thus, thanks to this new filiation, the presence of certain themes of research in the writings of Ibn al-Haytham, not only his work on the dioptre, the burning sphere and the spherical lens, is clarified; it authorizes what was not possible previously: to assess the distance covered by a generation of optical research—a distance so much more important, from the historical and the epistemological point of view, now that we are on the eve of one of the first revolutions in optics, if not in physics. Compared with the writings of the Greek and Arab mathematicians who preceded him, the optical work by Ibn al-Haytham presents at first glance two striking features: extension and reform. It will be concluded on a more careful examination that the first trait is the material trace of the second. In fact no one before Ibn al-Haytham had embraced so many domains in his research, collecting together fairly independent traditions: mathematical, philosophical, medical. The titles of his books serve moreover to illustrate this large spectrum: *The Light of the Moon*, *The Light of the Stars*, *The Rainbow and the Halo*, *Spherical Burning Mirrors*, *Parabolical Burning Mirrors*, *The Burning Sphere*, *The Shape of the Eclipse*, *The Formation of Shadow*s, *On Light*, as well as his *Book of Optics* translated into Latin in the twelfth century and studied and commented on in Arabic and Latin until the seventeenth century. Ibn al-Haytham therefore studied not only the traditional themes of optical research but also other new ones to cover finally the following areas: optics, catoptrics, dioptrics, physical optics, meteorological optics, burning mirrors, the burning sphere.

A more meticulous look reveals that, in the majority of these writings, Ibn al-Haytham pursued the realization of his programme to reform the discipline, which brought clearly to take up each different problem in turn. The founding action of this reform consisted in making clear the distinction, for the first time in the history of optics, between the conditions of propagation of light and the conditions of vision of objects. It led, on one hand, to providing physical support for the rules of propagation—it concerns a mathematically guaranteed analogy between a mechanical model of the movement of a solid ball thrown against an obstacle, and that of the light—and, on the other hand, to proceeding everywhere geometrically and by observation and experimentation. It led also to the definition of the concept of light ray and light bundle as a set of straight lines on which light propagates, rays independent from each other which propagate in a homogeneous region of space. These rays are not modified by other rays which propagate in the same region. Thanks to the concept of light bundle, Ibn al-Haytham was able to study the propagation and diffusion of light mathematically and experimentally. Optics no longer has the meaning that is assumed formerly: a geometry of perception. It includes henceforth two parts: a theory of vision, with which is also associated a physiology of the eye and a psychology of perception, and a theory of light, to which are linked geometrical optics and physical optics. Without doubt traces of the ancient optics are still detected: the survival of ancient terms, or

a tendency to pose the problem in relation to the subject of vision without that being really necessary. But these relics do not have to deceive: their effect is no longer the same, nor is their meaning. The organization of his *Book of Optics* reflects already the new situation. In it are books devoted in full to propagation—the third chapter of the first book and Books IV to VII; others deal with vision and related problems. This reform led, amongst other things, to the emergence of new problems, never previously posed, such as the famous "problem of Alhazen" on catoptrics, the examination of the spherical lens and the spherical dioptre, not only as burning instruments but also as optical instruments, in dioptrics; and to experimental control as a practice of investigation as well as the norm for proofs in optics and more generally in physics.

Let us follow now the realization of his reform in the *Book of Optics* and in other treatises. This book opens with a rejection and a reformulation. Ibn al-Haytham rejects straightaway all the variants of the doctrine on the visual ray, to ally himself with philosophers who defended an intromissionist doctrine on the form of visible objects. A fundamental difference remains nevertheless between him and the philosophers, such as his contemporary Avicenna: Ibn al-Haytham did not consider the forms perceived by the eyes as "totalities" which radiate from the visible object under the effect of light, but as reducible to their elements: from every point of the visible object radiate a ray towards the eye. The latter has become without soul, without πνεῦμα ὀπτικόν, a simple optical instrument. The whole problem was then to explain how the eye perceives the visible object with the aid of these rays emitted from every visible point.

After a short introductory chapter, Ibn al-Haytham devotes two successive chapters—the second and the third books of his *Book of Optics*—to the foundations of the new structure. In one, he defines the conditions for the possibility of vision, while the other is about the conditions for the possibility of light and its propagation. These conditions, which Ibn al-Haytham presents in the two cases as empirical notions, i.e. as resulting from an ordered observation or a controlled experiment, are effectively constraints on the elaboration of the theory of vision, and in this way on the new style of optics. The conditions for vision detailed by Ibn al-Haytham are six: the visible object must be luminous by itself or illuminated by another; it must be opposite to the eye, i.e. one can draw a straight line to the eye from each of its points; the medium that separates it from the eye must be transparent, without being cut into by any opaque obstacle; the visible object must be more opaque than this medium; it must be of a certain volume, in relation to the visual sharpness. These are the notions, writes Ibn al-Haytham, "without which vision cannot take place". These conditions, one cannot fail to notice, do not refer, as in the ancient optics, to those of light or its propagation. Of these, the most important, established by Ibn al-Haytham, are the following: light exists independently of vision and exterior to it; it moves with great speed and not instantaneously; it loses intensity as it moves away from the source; the light from a luminous source—substantial—and that from an illuminated object—second or accidental—propagate onto bodies which surround them, penetrate transparent media, and light up opaque bodies which in turn emit light; the light propagates from every point of the luminous or illuminated object in straight lines in transparent media and in all directions; these virtual straight lines along which light propagates form with it "the rays"; these lines can be parallel or cross one another, but the light does not mix in either case; the reflected or refracted light propagates along straight lines in particular directions. As can be noted, none of these notions relate to vision. Ibn al-Haytham completes them with other notions relative to colour. According to him, the colours exist independently from the light in opaque bodies, and as a consequence only light emitted by these bodies—second or accidental light—accompanies the colours which propagate then according to the same principles and laws as the light. As we have explained elsewhere, it is this

doctrine on colours which imposed on Ibn al-Haytham concessions to the philosophical tradition, obliging him to keep the language of "forms", already devoid of content when he only deals with light.[8]

A theory of vision must henceforth answer not only the six conditions of vision, but also the conditions of light and its propagation. Ibn al-Haytham devotes the rest of the first book of his *Book of Optics* and the two following books to the elaboration of this theory, where he takes up again the physiology of the eye and a psychology of perception as an integral part of this new intromissionist theory.

Three books of the *Book of Optics*—the fourth to the sixth—deal with catoptrics. This area, as ancient as the discipline itself, amply studied by Ptolemy in his *Optics*, has never been the object of so extensive a study as that by Ibn al-Haytham. Besides the three voluminous books of his *Book of Optics*, Ibn al-Haytham devotes other essays to it which complete them, on the subject of connected problems such as that of burning mirrors. Research into catoptrics by Ibn al-Haytham distinguishes itself, among other traits, by the introduction of physical ideas, both to explain the known ideas and to grasp new phenomena. It is in the course of this study that Ibn al-Haytham poses himself new questions, such as the problem that bears his name.

Let us consider some aspect of this research into catoptrics by Ibn al-Haytham. He restates the law of reflection, and explains it with the help of the mechanical model already mentioned. Then he studies this law for different mirrors: plane, spherical, cylindrical and conical. In each case, he applies himself above all to the determination of the tangent plane to the surface of the mirror at the point of incidence, in order to determine the plane perpendicular to this last plane, which includes the incident ray, the reflected ray and the normal to the point of incidence. Here as in his other studies, to prove these results experimentally, he conceives and builds an apparatus inspired by the one that Ptolemy constructed to study reflection, but more complicated and adaptable to every case. Ibn al-Haytham also studies the image of an object and its position in the different mirrors. He applies himself to a whole class of problems: the determination of the incidence of a given reflection in the different mirrors and conversely. He also poses for the different mirrors the problem which his name is associated with: given any two points in front of a mirror, how does one determine on the surface of the mirror a point such that the straight line which joins the point to one of the two given points is the incident ray, whilst the straight line that joins this point to the other given point is the reflected ray. This problem, which rapidly becomes more complicated, has been solved by Ibn al-Haytham.

Ibn al-Haytham pursues this catoptric research in other essays, some of which are later than the *Book of Optics*, such as *Spherical Burning Mirrors*.[9] It is in this essay of a particular interest that Ibn al-Haytham discovers the longitudinal spherical aberration; it is also in this text that he proves the following proposition:

On a sphere of centre E let there be a zone surrounded by two circles of axis EB; let IJ be the generator arc of this zone, and D its midpoint. Ibn al-Haytham has shown in two previous propositions that to each of the two circles is associated a point of the axis towards which the incident rays parallel to the axis reflect on this circle. He shows here that all the rays reflected on the zone meet the segment thus defined: if GD is the medium ray of the zone, the point H is associated with D, and the segment is on either side of H. The length of this segment depends on the arc IJ (Fig. 2.3).

8 See Rashed [4], pp. 271–298.

9 Ibn al-Haytham [1].

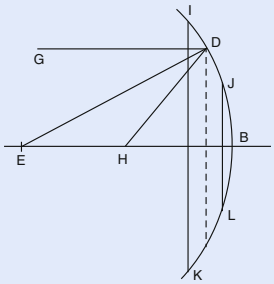


Fig. 2.3 Ibn al-Hayhtam illustration of the longitudinal spherical aberration

The seventh and last book of the *Book of Optics* by Ibn al-Haytham is devoted to dioptrics. In the same way as he did for catoptrics, Ibn al-Haytham inserts in this book the elements of a physical—mechanical—explanation of refraction. Moreover, his book is completed by his essays, such as his treatise on the *Burning Sphere* or his *Discourse on Light*, where he comes back to the notion about the medium, following Ibn Sahl.

In this seventh book of the *Book of Optics*, Ibn al-Haytham starts by taking on the two qualitative laws of refraction, and several quantitative rules, all controlled experimentally with the help of an apparatus that he conceives and builds as in the previous case. The two quantitative laws known by his predecessors, Ptolemy and Ibn Sahl, can be expressed as follows: (1) the incident ray, the normal at the point of refraction and the refracted ray are in the same plane; the refracted ray approaches (or moves away from) the normal if the light passes from a less (respectively more) refractive medium to a more (respectively less) refractive medium; (2) the principle of the inverse return.

But, instead of following the way opened by Ibn Sahl through his discovery of the law of Snellius, Ibn al-Haytham returns to the ratios of angles and establishes his quantitative rules.

1. The angles of deviation vary in direct proportion to the angles of incidence: if in medium n_1 one takes $i' > i$, one will have, in medium n_2, $d' > d$ (i is the angle of incidence, r the angle of refraction and d the angle of deviation; $d = |i - r|$).
2. If the angle of incidence increases by a certain amount, the angle of deviation increases by a smaller quantity: if $i' > i - I$ and $d' > d$, one will have $d' - d < i' - i$.
3. The angle of refraction increases in proportion to the angle of incidence: if $i' > i$, one will have $r' > r$.
4. If the light penetrates from a less refractive medium into a more refractive medium, $n_1 < n_2$, one has $d < \frac{1}{2}i$; in the opposite path, one has $d < \frac{(i+d)}{2}$, and one will have $2i > r$.
5. Ibn al-Haytham takes up again the rules stated by Ibn Sahl in his book on *The Celestial Sphere*; he affirms that, if the light penetrates from a medium n_1, with the same angle of incidence, into two different media n_2 and n_3, then the angle of deviation is different for each of these media because of the difference in opaqueness. If, for example, n_3 is more opaque than n_2, then the angle of deviation will be larger in n_3 than in n_2. Conversely, if n_1 is more opaque than n_2, and n_2 more opaque than n_3, the angle of deviation will be larger in n_3 than in n_2.

Contrary to what Ibn al-Haytham believes, these quantitative rules are not all valid in a general sense. But to his credit all are provable within the limits of the

experimental conditions he effectively envisaged in his *Book of Optics*; the media are air, water and glass, with angles of incidence which do not go above 80°.[10]

Ibn al-Haytham devotes a substantial part of the seventh book to the study of the image of an object by refraction, notably if the surface of separation of the two media is either plane or spherical. It is in the course of this study that he settles on the spherical dioptre and the spherical lens, following thus in some way the research by Ibn Sahl, but modifying it considerably; this study of the dioptre and the lens appears in effect in the chapter devoted to the problem of the image, and is not separated from the problem of vision. For the dioptre, Ibn al-Haytham considers two cases, depending on whether the source—punctual and at a finite distance—is found on the concave or convex side of the spherical surface of the dioptre.

Ibn al-Haytham studies the spherical lens, giving particular attention to the image that it gives of an object. He restricts himself nevertheless to the examination of only one case, when the eye and the object are on the same diameter. Put another way, he studies the image through a spherical lens of an object placed in a particular position on the diameter passing through the eye. His procedure is not without similarities to that of Ibn Sahl when he studied the biconvex hyperbolic lens. Ibn al-Haytham considers two dioptres separately, and applies the results obtained previously. It is in the course of his study of the spherical lens that Ibn al-Haytham returns to the spherical aberration of a point at a finite distance in the case of the dioptre, in order to study the image of a segment which is a portion of the segment defined by the spherical aberration.

In his treatise on the *Burning Sphere*, Ibn al-Haytham explains and refines certain results on the spherical lens which he had already obtained in his *Book of Optics*. However, he returns to the question of the burning by means of that lens. It is in this treatise that we encounter the first deliberate study of spherical aberration for parallel rays falling on a glass sphere and undergoing two refractions. In the course of this study, Ibn al-Haytham uses numerical data given in the *Optics* by Ptolemy for the two angles of incidence 40° and 50°, and, to explain this phenomenon of focusing of light propagated along trajectories parallel to the diameter of the sphere, he returns to angular values instead of applying what is called the law of Snellius.

In this treatise on the *Burning Sphere*, as in the seventh book of his *Book of Optics* or in other writings on dioptrics, Ibn al-Haytham exposes his research in a somewhat paradoxical way: while he takes a lot of care to invent, fashion and describe some experimental devices that are advanced for this age, allowing the determination of numerical values, in most cases he avoids giving these values. When he does give them, as in the treatise on the *Burning Sphere*, it is with economy and circumspection. For this attitude, already noted, at least two reasons can perhaps be found. The first is in the style of the scientific practice itself: quantitative description does not yet seem to be a compelling norm. The second is no doubt linked: the experimental devices can only give approximate values. It is for this reason that Ibn al-Haytham took into account the values which he had borrowed from the *Optics* by Ptolemy.

This book on the *Burning Sphere* is undoubtedly one of the summits of research in classical optics. Kamāl al-Dīn al-Fārisī (d. 1319) was able to put this book to work in order to explain for the first time the rainbow and the hallo. In this book, Ibn al-Haytham returns to the problem of combustion with the help of a spherical lens. Here then, is a text that enables us to follow the evolution of Ibn al-Haytham's thought on spherical lenses, by examining how he takes up the problem raised by his predecessor Ibn Sahl: to cause combustion by refraction,

10 See Rashed [3].

with the help of a lens. For Ibn al-Haytham, this research is an integral part of optics.

He begins this book by proving several propositions two of which are particularly important:

1. $\frac{i}{4} < d < \frac{i}{2}$ (i, angle of incidence in the glass; d, angle of deviation)
2. Let α and β be two arcs of a circle; we suppose that $\alpha > \beta$; $\alpha = \alpha_1 + \alpha_2$ and $\beta = \beta_1 + \beta_2$, such that

$$\frac{\alpha_2}{\alpha_1} = \frac{\beta_2}{\beta_1} = k < 1$$

and

$$\alpha_1 < \frac{\pi}{2}$$

Then $\dfrac{\sin\beta_1}{\sin\beta_2} > \dfrac{\sin\alpha_1}{\sin\alpha_2}$.

With the help of these two propositions, as well as his rules of refraction, Ibn al-Haytham studies the propagation of a bundle of parallel rays falling upon a glass or crystal sphere. Let us sketch how he proceeds.

In a first proposition, he shows that all parallel rays falling on the sphere with one and the same angle of incidence converge, after two refractions, towards one and the same point of the diameter which is parallel to the ray. This point is the focus associated with incidence i.

Thus, he considers a ray (HN) parallel to the diameter AC, falling upon the sphere at M. The refracted ray corresponding to it meets the sphere at B, and meets AC at point S. Point S is the focus associated with incidence i, and it belongs to the segment $[CK]$ where K is the intersection of MB with AC (Fig. 2.4).

In a second proposition, he proves that the total deviation is twice each of the deviations: $D = 2d$.

He proves then that a given point S, beyond C on the diameter, can be obtained only from a single point M; that is to say, S corresponds to a single incidence.

In a third proposition, he proves that the two incidences i and i', correspond two distinct points S and S'.

In a fourth proposition, he proves: for $i > i'$, we have S and S' such that $CS' > CS$. Therefore, when i increases, CS decreases. To a given point S, therefore, there corresponds one single incidence i.

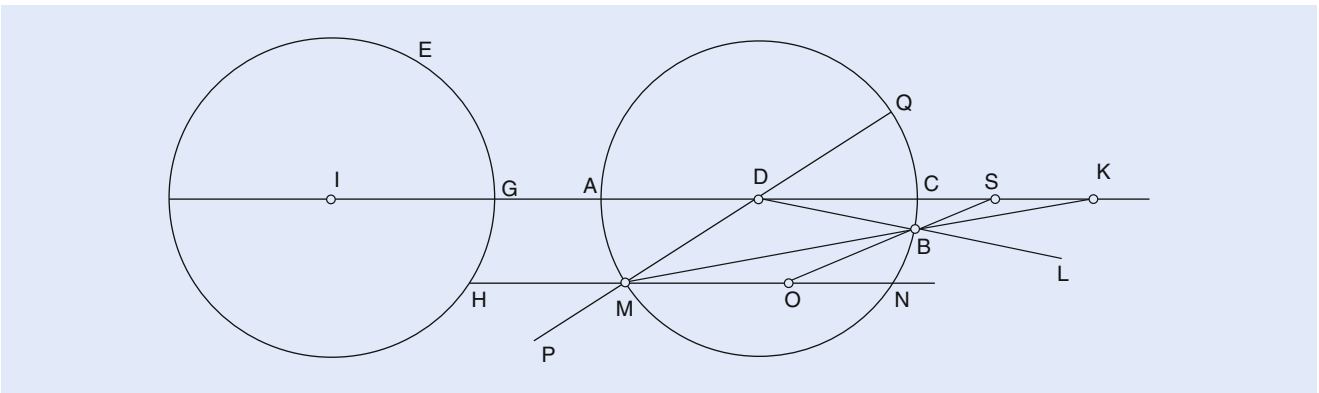


Fig. 2.4 Illustration of spherical aberration in glass (crystal) spheres

Ibn al-Haytham proposes to determine the extremities of the segment on which the points *S* are located. With this in view, he studies the positions of *B*—the point of the second refraction—when the angle of incidence varies. As far as we know, this is the first deliberate study of spherical aberration for parallel rays which fall on a glass sphere and undergo two refractions.[11]

2.4 Conclusion

Let us stop at this point on spherical aberration, to conclude.

With Ibn al-Haytham, one result has been definitively obtained: the half century which separates him from Ibn Sahl should be counted among the distinctive moments in the history of optics: dioptrics appears to have extended its domain of validity and, by its very progress, to have changed its orientation. With Ibn al-Haytham, the conception of dioptrics as a geometry of lenses has become outdated. Here again, in his own words, we must combine mathematics and physics in order to study dioptres and lenses, whether burning or not. The mathematization could only be achieved with Ibn al-Haytham because he separated the study of the natural phenomenon of light from vision and sight. The step taken suggests already that the domain carved out by Ibn Sahl was not long-lived and wound up, 50 years later, exploding under the assault of the mathematician and physicist Ibn al-Haytham. In optics as in astronomy the research programme of Ibn al-Haytham is the same: mathematize the discipline and combine this mathematization with the ideas of the natural phenomena.

References

1. Ibn al-Haytham. Fī al-marāyā al-muḥriqa bi-al-dawā'ir, MS Berlin, Oct 2970/7, fols 66r-73v
2. Ibn al-Haytham (1989) The Optics of Ibn al-Haytham, I, Books I–III (trans: Sabra AI). London
3. Rashed R (1968) Le Discours de la lumière d'Ibn al-Haytham (Alhazen). Revue d'Histoire des Sciences 21(1968):197–224, repr. in Optique et Mathématiques: Recherches sur l'histoire de la pensée scientifique en arabe, Variorum reprints, Aldershot
4. Rashed R (1992) Optique géométrique et doctrine optique chez Ibn al-Haytham. In: Optique et Mathématiques: Recherches sur l'histoire de la pensée scientifique en arabe. Variorum reprints, Aldershot
5. Rashed R (1997) Œuvres philosophiques et scientifiques d'al-Kindī. Vol. I: L'Optique et la Catoptrique d'al-Kindī. E.J. Brill, Leiden; Arabic translation: '*Ilm al-manāẓir wa-'ilm in'ikās*

11 See Rashed [7], p. 164.

al-ḍaw', Silsilat Tārīkh al-'ulūm 'inda al-'Arab 6, Beirut, Markaz Dirāsat al-Waḥda al-'Arabiyya
6. Rashed R (2000) Les Catoptriciens grecs. I: Les Miroirs ardents, édition, traduction et commentaire, Collection des Universités de France, publiée sous le patronage de l'Association Guillaume Budé. Les Belles Lettres, Paris
7. Rashed R (2005) Geometry and Dioptrics in Classical Islam. Al-Furqān, London
8. Rashed R (2014) Ibn al-Haytham. New Spherical Geometry and Astronomy. A History of Arabic Sciences and Mathematics, vol. 4, Culture and Civilization in the Middle East, London, Centre for Arab Unity Studies, Routledge

Ultrafast Phenomena and the Invisible World

Contents

Ultrafast Light and Electrons: Imaging the Invisible

Ahmed H. Zewail

A.H. Zewail (✉)
Physical Biology Center for Ultrafast Science and Technology, Arthur Amos Noyes Laboratory for Chemical Physics, California Institute of Technology, Pasadena, CA 91125, USA
e-mail: zewail@caltech.edu

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_3

3

3.1 Origins

The ever-increasing progress made by humans in making the very small and the very distant visible and tangible is truly remarkable. The human eye has spatial and temporal resolutions that are limited to about 100 μm and a fraction of a second, respectively. Today we are aided by tools that enable the visualization of objects that are below a nanometer in size and that move in femtoseconds ([98] and the references therein). This chapter, which is based on review articles by the author [94, 99–101], some commentaries [83–85, 87], and a book [103], provides a road map for the evolutionary and revolutionary developments in the fields of ultrafast light and electrons used to image the invisibles of matter.

How did it all begin? Surely, the power of light for observation has been with humans since their creation. Stretching back over six millennia, one finds its connection to the science of time clocking [95] (first in calendars) and to the mighty monotheistic faiths and rituals (Fig. 3.1). Naturally, the philosophers of the past must have been baffled by the question: what is light and what gives rise to the associated optical phenomena?

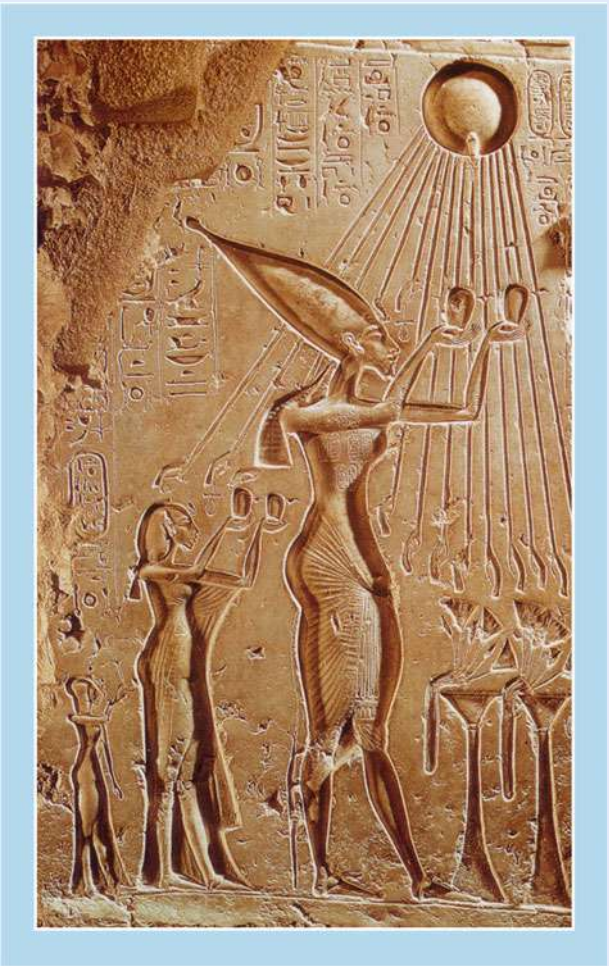


Fig. 3.1 The significance of the light–life interaction as perceived more than three millennia ago, at the time of Akhenaton and Nefertiti. Note the light's "ray diagram" from a spherical source, the Sun. Adapted from Zewail [96]

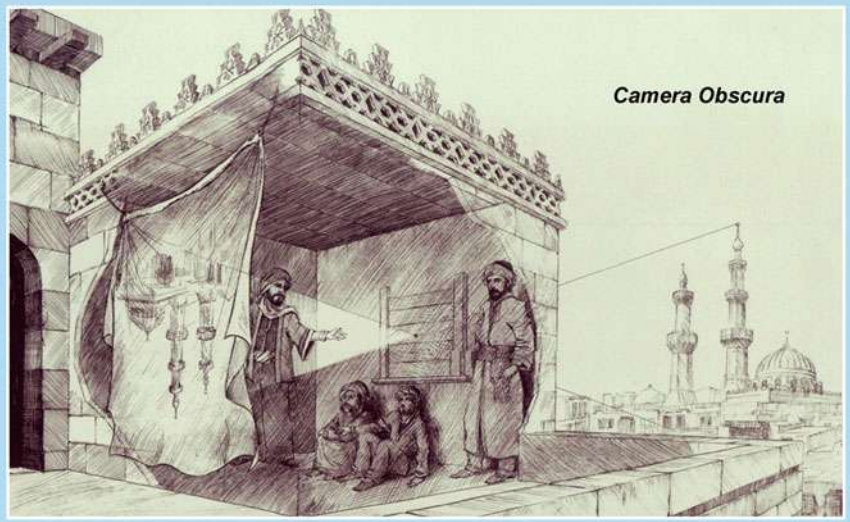


Fig. 3.2 The concept of the *camera obscura* as perceived a 1000 years ago by Alhazen (Ibn al-Haytham), who coined the term and experimented with the light rays (see Sec. 3.1). Note the formation of the inverted image through a ray diagram. Adapted from Al-Hassani et al. [1]

A leading contribution to this endeavor was made by the Arab polymath Alhazen (Ibn al-Haytham; AD 965–1040) nearly a millennium ago. He is recognized for his quantitative experimentation and thoughts on light reflection and refraction, and is also credited with explaining correctly the mechanism of vision, prior to the contributions of Kepler, Descartes, Da Vinci, Snell, and Newton. But of relevance to our topic is his conceptual analysis of the *camera obscura*, the "dark chamber," which aroused the photographic interests of J. W. Strutt (later known as Lord Rayleigh) in the 1890s [81]. Alhazen's idea that light must travel along straight lines and that the object is inverted in the image plane is no different from the modern picture of ray diagrams taught in optics today (Fig. 3.2). His brilliant work was published in the *Book of Optics* or, in Arabic, *Kitab al-Manazir*.

3.2 Optical Microscopy and the Phenomenon of Interference

In the fourteenth and fifteenth centuries, the art of grinding lenses was perfected in Europe, and the idea of optical microscopy was developed. In 1665, Robert Hooke (the scientist who coined the word "cell") published his studies in *Micrographia* ([39]; Fig. 3.3), and among these studies was a description of plants, feathers, as well as cork cells and their ability to float in water. Contemporaneously, Anton van Leeuwenhoek used a simple, one-lens microscope to examine blood, insects, and other objects, and was the first to visualize bacteria, among other microscopic objects.

More than a 100 years later, an experiment by the physicist, physician, and Egyptologist, Thomas Young, demonstrated the interference of light, an experiment that revolutionized our view of the nature of light. His double-slit experiment of 1801 performed at the Royal Institution of Great Britain led to the rethinking of Newton's corpuscular theory of light. Of relevance here is the phenomenon of diffraction due to interferences of waves (coherence). Much later, such diffraction

Fig. 3.3 Microscopy time line, from *camera obscura* to 3D electron microscopes. 4D ultrafast electron microscopy and diffraction were developed a decade ago (see Sec. 3.7). The top inset is the frontispiece of Hooke's [39] *Micrographia* published by the Royal Society of London. In the frontispiece to Hevelius's *Selenographia* (bottom inset), Ibn al-Haytham represents *Ratione* (the use of reason) with his geometrical proof and Galileo represents Sensu (the use of the senses) with his telescope. The two scientists hold the book's title page between them, suggesting a harmony between the methods [72, 80]. Adapted from Zewail and Thomas [98]

was found to yield the (microscopic) interatomic distances characteristic of molecular and crystal structures, as discovered in 1912 by von Laue and elucidated later that year by W. L. Bragg.

Resolution in microscopic imaging was brought to a whole new level by two major developments in optical microscopy. In 1878, Ernst Abbe formulated a mathematical theory correlating resolution to the wavelength of light (beyond what we now designate the empirical Rayleigh criterion for incoherent sources), and hence the optimum parameters for achieving higher resolution. At the

beginning of the twentieth century, Richard Zsigmondy, by extending the work of Faraday and Tyndall, developed the "ultramicroscope" to study colloidal particles; for this work, he received the Nobel Prize in Chemistry in 1925. Then came the penetrating developments in the 1930s by Frits Zernike, who introduced the phase-contrast concept in microscopy; he, too, received the Nobel Prize, in Physics, in 1953. It was understood that the spatial resolution of optical microscopes was limited by the wavelength of the visible light used. Recently, optical techniques have led to considerable improvement in the spatial resolution, and the 2014 Nobel Prize in Chemistry was awarded to Eric Betzig, Stefan Hell, and William Moerner for reaching the spatial resolution beyond the diffraction limit.

3.3 The Temporal Resolution: From Visible to Invisible Objects

In 1872 railroad magnate Leland Stanford wagered $25,000 that a galloping horse, at some point in stride, lifts all four hooves off the ground. To prove it, Stanford employed English photographer Eadweard Muybridge. After many attempts, Muybridge developed a camera shutter that opened and closed for only two thousandths of a second, enabling him to capture on film a horse flying through the air (Fig. 3.4). During the past century, all scientific disciplines from astrophysics to zoology have exploited "high-speed" photography to understand motion of objects and animals that are quicker than the eye can follow.

The time resolution, or shutter speed, needed to photograph the ultrafast motions of atoms and molecules is beyond any conventional scale. When a molecule breaks apart into fragments or when it combines with another to form a new molecule, the chemical bonds between atoms break or form in less than a trillionth of a second, or one picosecond. Scientists have hoped to observe molecular motions in real time and to witness the birth of molecules: the instant at which the fate of the molecular reaction is decided and the final products are determined. Like Muybridge, they needed to develop a shutter, but it had to work ten billion times faster than that of Muybridge.

In the 1980s, our research group at the California Institute of Technology developed new techniques to observe the dynamics of molecules in real time [95]. We integrated our system of advanced lasers with molecular beams, rays of isolated molecules, to a point at which we were able to record the motions of molecules as they form and break chemical bonds on their energy landscapes (Fig. 3.4). The chemical reaction can now be seen as it proceeds from reactants through transition states, the ephemeral structures between reactants and products, and finally to products—chemistry as it happens!

Because transition states exist for less than a trillionth of a second, the time resolution should be shorter—a few quadrillionths of a second, or a few femtoseconds (1 fs is equal to 10^{-15} s). A femtosecond is to a second what a second is to 32 million years. Furthermore, whereas in 1 s light travels nearly 300,000 km—almost the distance between the Earth and the Moon—in 1 fs light travels 0.3 μm, about the diameter of the smallest bacterium.

Over half a century ago, techniques were introduced to study the so-called chemical intermediates using fast kinetics. Ronald Norrish and George Porter of the University of Cambridge and Manfred Eigen of the Max Planck Institute for Physical Chemistry were able to resolve chemical events that lasted about a thousandth of a second, and in some cases a millionth of a second. This time scale was ideal for the intermediates studied, but too long for capturing the transition states—by orders of magnitude for what was needed. Over the millisecond—or even the nano/picosecond—time scale, the transition states are not in the picture, and hence not much is known about the reaction scenery.

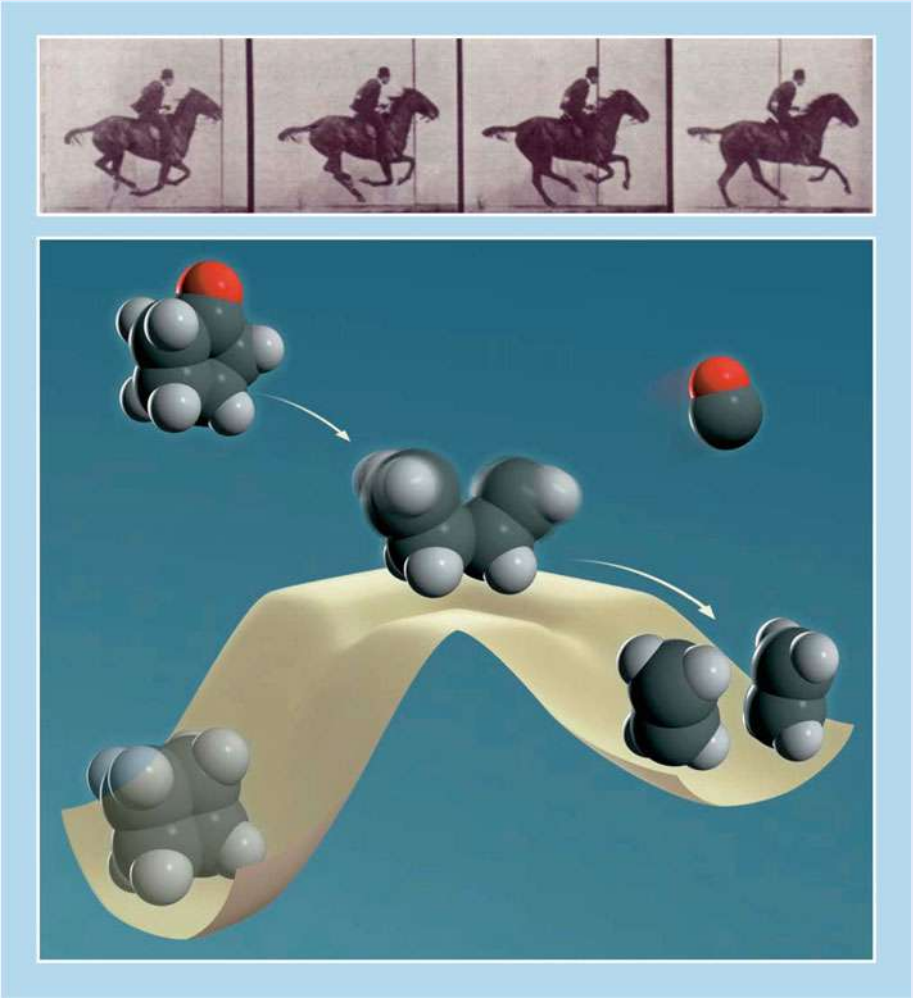


Fig. 3.4 Ultrafast and fast motions. The snapshots of a galloping horse in the upper panel were taken in 1887 by Eadweard Muybridge. He photographed a single horse and the shutter speed was 0.002 s per frame. In the molecular case (*femtochemistry*), a ten-orders-of-magnitude improvement in the temporal resolution was required, and methods for synchronization of billions of molecules had to be developed. The yellow surface in the lower panel represents the energy landscape for the transitory journey of the chemical reaction. Adapted from Zewail [94, 95]

Chemist Sture Forsén of Lund University came up with an insightful analogy that illustrates the importance of understanding transitory stages in the dynamics. He compared researchers to a theater audience watching a drastically shortened version of a classical drama. The audience is shown only the opening scenes of, say, *Hamlet* and its finale. Forsén writes, "*The main characters are introduced, then the curtain falls for change of scenery and, as it rises again, we see on the scene floor a considerable number of 'dead' bodies and a few survivors. Not an easy task for the inexperienced to unravel what actually took place in between*" [25].

The principles involved in the ultrafast molecular "camera" [42] have some similarity to those applied by Muybridge. The key to his work was a special camera shutter that exposed a film for only 0.002 s. To set up the experiment, Muybridge spaced 12 of these cameras half a meter apart alongside a horse track. For each camera he stretched a string across the track to a mechanism that would trigger the shutter when a horse broke through the string. With this system, Muybridge attained a resolution in each picture of about two centimeters, assuming the horse was galloping at a speed of about 10 m/s. (The resolution, or definition, is

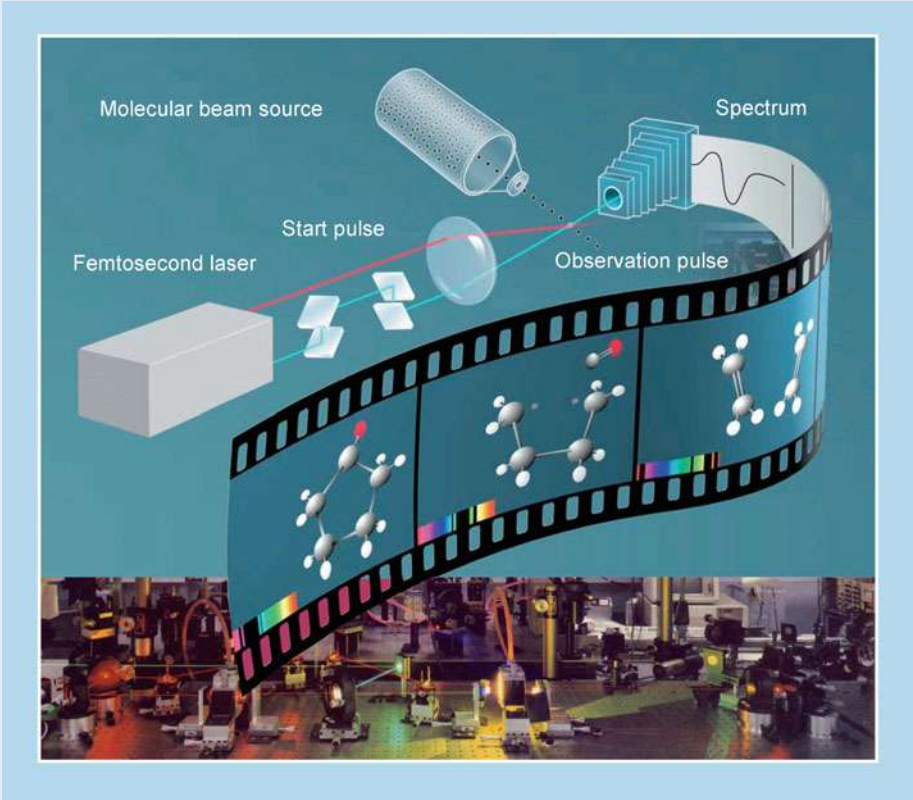


Fig. 3.5 "The fastest camera in the world" [42] records what happens during a molecular transformation by initiating the reaction with a femtosecond laser pulse ("start pulse," *red*). A short time later a second pulse ("observation pulse," *blue*) takes a "picture" of the reacting molecule(s). By successively delaying the observation pulse in relation to the start pulse a "film" is obtained of the course of the reaction. With this first ultrafast "camera" built at Caltech, the ephemeral *transition states* were identified and characterized. The bottom inset shows part of the camera. It is a complex array of lasers, mirrors, lenses, prisms, molecular beams, detection equipment, and more. Adapted from Zewail [95]

simply the velocity of the motion multiplied by the exposure time.) The speed of the motion divided by the distance between cameras equals the number of frames per second—20 in this case. The motion within a picture becomes sharper as the shutter speed increases. The resolution of the motion improves as the distance between the cameras decreases.

Two aspects are relevant to the femtosecond, molecular camera (Fig. 3.5). First, a continuous motion is broken up into a series of snapshots or frames. Thus, one can slow down a fast motion as much as one likes so that the eye can see it. Second, both methods must produce enough frames in rapid succession so that the frames can be reassembled to give the illusion of a continuous motion. The change in position of an object from one frame to the next should be gradual, and at least 30 frames should be taken to provide 1 s of the animation. In our case—the femtosecond, molecular camera—the definition of the frame and the number of frames per second must be adjusted to resolve the elementary nuclear motions of reactions and, most importantly, the ephemeral transition states. The frame definition must be shorter than 0.l nm. Because the speed of the molecular motion is typically 1 km/s, the shutter resolution must be in a time range of better than 100 fs.

Experimentally, we utilize the femtosecond pulses in the configuration shown in Fig. 3.5. The first femtosecond pulse, called the pump pulse, hits the molecule to initialize the reaction and set the experimental clock "at zero." The second laser pulse, called the probe pulse, arrives several femtoseconds later and records a

snapshot of the reaction at that particular instant. Like the cameras in Muybridge's experiment, the femtosecond, molecular camera records successive images at different times in order to obtain information about different stages of the reaction. To produce time delays between the pump and the probe pulses, we initially tune the optical system so that both pulses reach the specimen at the same time. We then divert the probe pulse so that it travels a longer distance than does the pump pulse before it reaches the specimen (Fig. 3.5). If the probe pulse travels one micrometer farther than the pump pulse, it will be delayed 3.33 fs, because light travels at 300,000 km/s. Accordingly, pulses that are separated by distances of 1–100 μm resolve the motion during 3.33–333 fs periods. A shutter speed of a few femtoseconds is beyond the capability of any camera based on mechanical or electrical devices. When the probe pulse hits the molecule, it does not then transmit an image to a detector; see below. Instead the probe pulse interacts with the molecule, and then the molecule emits, or absorbs, a spectrum of light (Fig. 3.6) or changes its mass.

When my colleague and friend Richard Bernstein of the University of California at Los Angeles learned about the femtosecond, molecular camera, he was very enthusiastic about the development, and we discussed the exciting possibilities created by the technique. At his house in Santa Monica, the idea of naming the emerging field of research *femtochemistry* was born. The field has now matured in many laboratories around the globe, and it encompasses applications in chemistry, biology, and materials science.

3.4 Electron Microscopy: Time-Averaged Imaging

Just before the dawn of the twentieth century, in 1897, electrons, or the *corpuscles* of J. J. Thomson, were discovered, but they were not conceived as imaging rays until Louis de Broglie formulated the concept of particle–wave duality in 1924. The duality character of an electron, which is quantified in the relationship $\lambda_{\text{deBroglie}} = h/p$, where h is Planck's constant and p is the momentum of the particle, suggested the possibility of achieving waves of picometer wavelength, which became essential to the understanding of diffraction and imaging. The first experimental evidence of the wave character of the electron was established in 1927 by Davisson and Germer (diffraction from a nickel surface) and, independently, by G. P. Thomson (the son of J. J. Thomson), who, with Reid, observed diffraction of electrons penetrating a thin foil. Around the same time, in 1923, Dirac postulated the concept of "single-particle interference."

With these concepts in the background, Knoll and Ruska [46] invented the electron microscope (EM), in the transmission mode—TEM, using accelerated electrons. Initially the resolution was close to that of optical microscopy, but, as discussed below, it now reaches the atomic scale! Many contributions (Fig. 3.3) to this field have laid the foundation ([56] and the references therein, [77, 79]) for advances of the fundamentals of microscopy and for recent studies of electron interferometry [88, 89]. A comprehensive overview is given by Zewail and Thomas [98].

3.5 2D Imaging and Visualization of Atoms

The first images of individual atoms were obtained in 1951 by Müller [64, 86, 90], who introduced the technique of field-ion microscopy to visualize them at fine tips of metals and alloys, and to detect vacancies and atomic steps and kinks at surfaces. With the invention of field-emission sources and scanning TEM, pioneered in 1970 by Crewe, isolated heavy atoms became readily visible [15, 82]. (The scanning

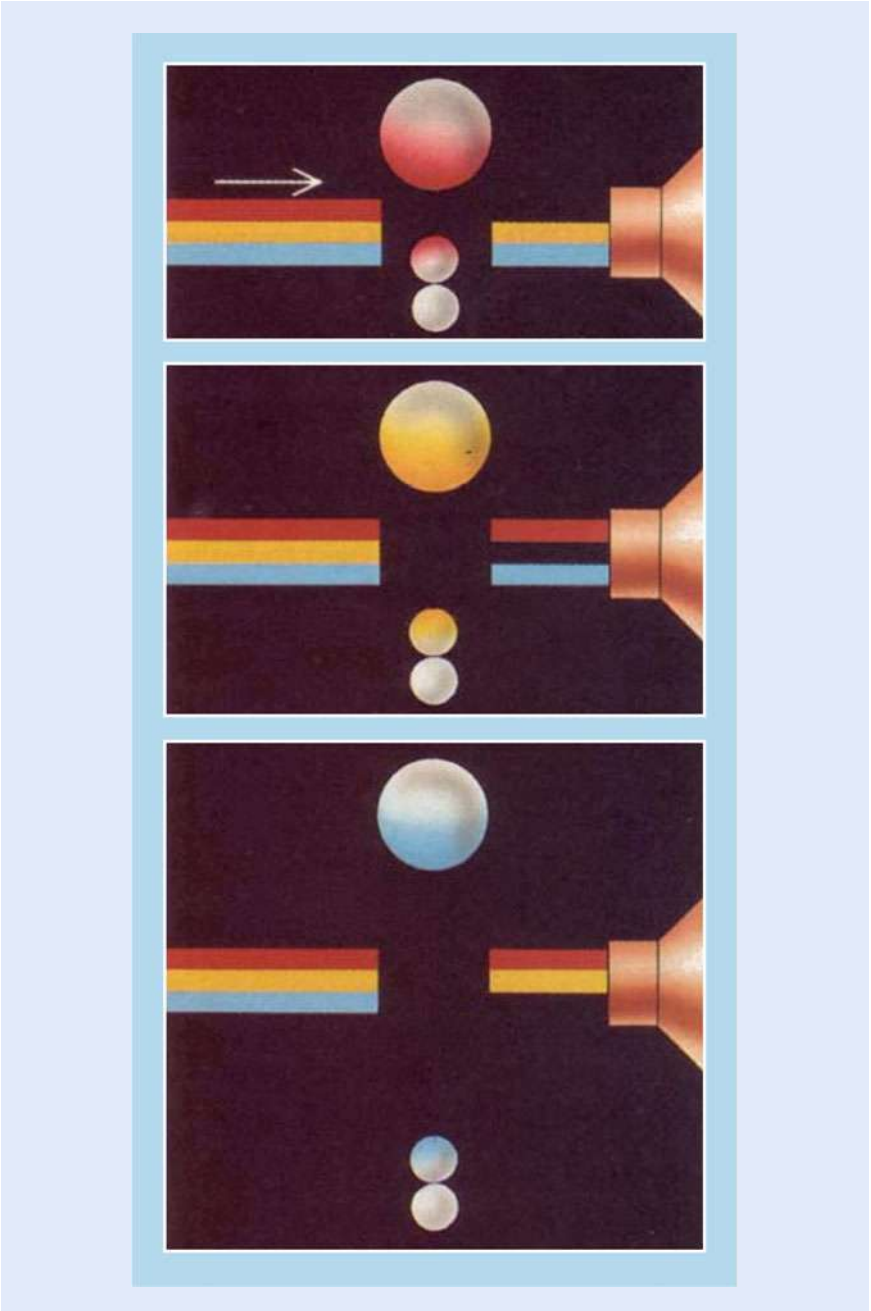


Fig. 3.6 Femtosecond spectroscopy of transient species. A given state of a molecular system can be identified by the light that the molecule absorbs. When atoms in a molecule are relatively close together, they tend to absorb long wavelengths of light (*red*, for example). When the atoms are farther apart, they tend to absorb short wavelengths of light (*blue*, for instance). The change in the spectrum is the fingerprint of the atoms in motion. Adapted from Zewail [94]

tunneling microscope was developed in the 1980s and made possible atomic-scale images of conducting surfaces.) Today, with aberration-corrected microscopes, imaging has reached a resolution of less than an ångström [65]. This history would be incomplete if I did not mention that the totality of technical developments and applications in the investigations of inorganic and organic materials have benefited enormously from the contributions of many other scientists, and for more details I refer the reader to the books by Cowley [14], Humphreys [41], Gai and Boyes [28], Spence [78], and Hawkes and Spence [34], and the most recent papers by Hawkes [33] and Howie [40].

3.6 The Third Dimension and Biological Imaging

Biological EM has been transformed by several major advances, including electron crystallography, single-particle tomography, and cryo-microscopy, aided by large-scale computational processing. Beginning with the 1968 electron crystallography work of DeRosier and Klug, see [45], 3D density maps became retrievable from EM images and diffraction patterns. Landmark experiments revealing the high-resolution structure from 2D crystals, single-particle 3D cryo-EM images of different but identical particles (6 Å resolution), and 3D cryo-EM images of the same particle (tomography with 6 Å resolution) represent the impressive progress made. Recently, another milestone in EM structural determination has been reported [2, 51]. Cryo-EM is giving us the first glimpse of mitochondrial ribosome at near-atomic resolution, and, as importantly, without the need for protein crystallization or extensive protocols of purification. In this case, the biological structure is massive, being three megadalton in content, and the subunit has a 39 protein complex which is clearly critical for the energy-producing function of the organelle. Using direct electron detection, a method we involved in the first ultrafast electron diffraction (UED) experiments [18, 91], the spatial resolution reached was 3.2 Å. With the recent structural advances made by the EM groups at the University of California, San Francisco (ion channels), Max Planck Institute of Biophysics, Frankfurt am Main (hydrogenase), and others, it is clear that EM is leading the way in the determination of macromolecular (and noncrystalline!) structures; the highlight by Kühlbrandt [51] provides the relevant references. The determined structures, however, represent an average over time.

With these methods, the first membrane protein structure was determined, the first high-resolution density maps for the protein shell of an icosahedral virus were obtained, and the imaging of whole cells was accomplished. Minimizing radiation damage by embedding the biological macromolecules and machines in vitreous ice affords a non-invasive, high-resolution imaging technique for visualizing the 3D organization of eukaryotic cells, with their dynamic organelles, cytoskeletal structure, and molecular machines in an unperturbed context, with an unprecedented resolution. I refer the reader to the papers by Henderson [35], Sali et al. [73], Crowther [16], and Glaeser [30], and the books by Glaeser et al. [31] and by Frank [27]. The Nobel paper by Roger Kornberg [48] on RNA polymerase II (pol II) transcription machinery is a must for reading. Recently, Henderson commented in a *Nature* article on the overzealous claims made by some in the X-ray laser community, emphasizing the unique advantages of electron microscopy and its cryo-techniques for biological imaging [36].

3.7 4D Ultrafast Electron Microscopy

Whereas in all of the above methods the processes of imaging, diffraction, and chemical analysis have been conducted in a *static* (time-averaged) manner, with the advent of femtosecond light pulses it has now become possible to unite the temporal domain with the (3D) spatial one, thereby creating 4D electron micrographs (Fig. 3.7; [99, 102]); the new approach is termed *4D ultrafast electron microscopy* or—for short—4D UEM. This development owes its success to the advancement of the concept of coherent *single-electron imaging* [99], with the electron packets being liberated from a photocathode using femtosecond optical pulses. In such a mode of electron imaging, the repulsion between electrons is negligible, and thus atomic-scale spatiotemporal resolution can be achieved. Atomic motions, phase transitions, mechanical movements, and the nature of fields at interfaces are examples of phenomena that can be charted in

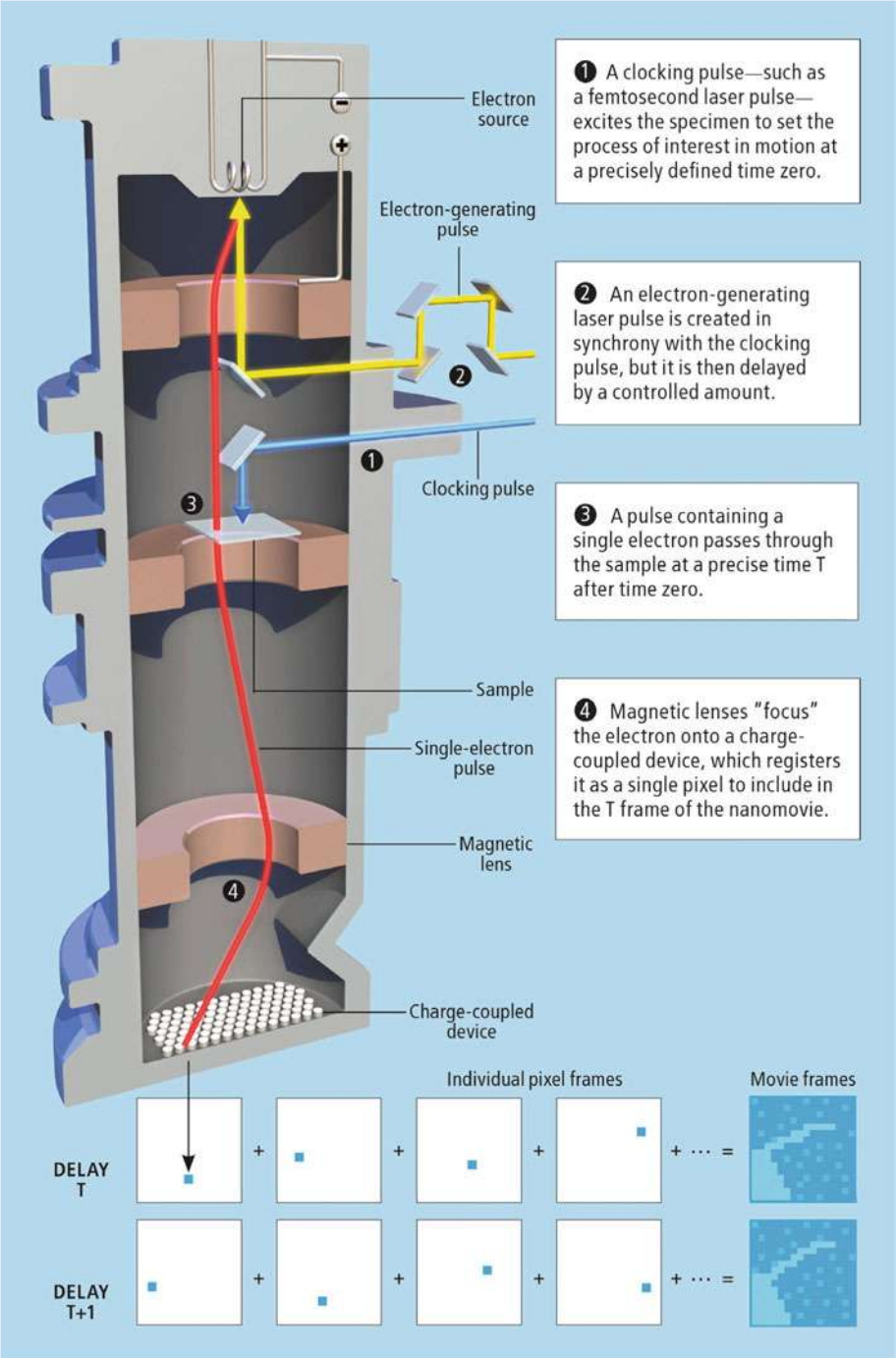


Fig. 3.7 Concept of single-electron 4D UEM. A standard electron microscope records still images of a nanoscopic sample by sending a beam of electrons through the sample and focusing it onto a detector. By employing single-electron pulses, a 4D electron microscope produces movie frames representing time steps as short as femtoseconds (10^{-15} s). Each frame of the nanomovie is built up by repeating this process thousands of times with the same delay and combining all the pixels from the individual shots. The microscope may also be used in other modes, e.g., with one many-electron pulse per frame, depending on the kind of movie to be obtained. The single-electron mode produces the finest spatial resolution and captures the shortest time spans in each frame. Adapted from Zewail [101]

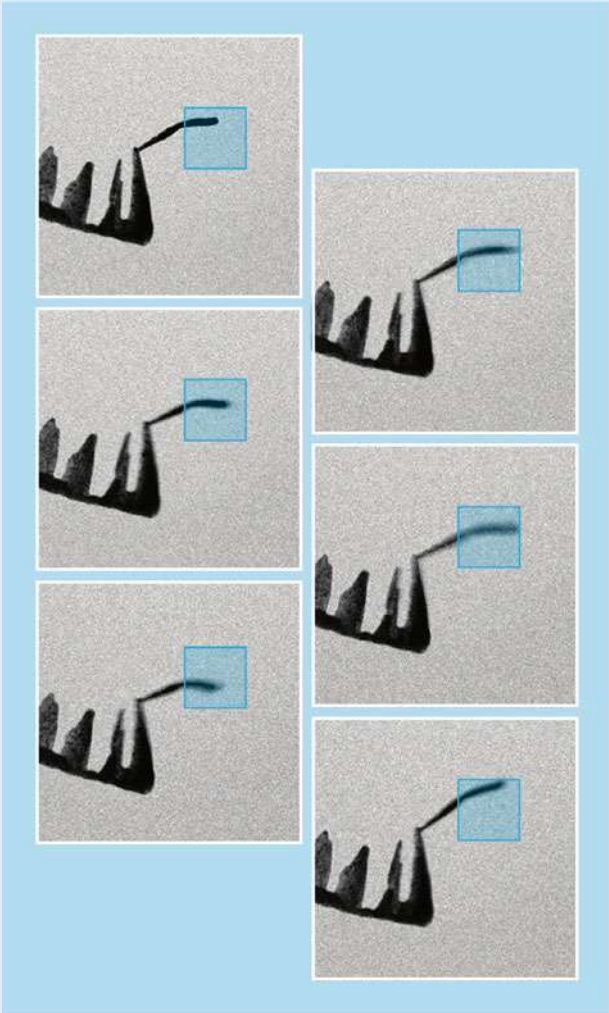


Fig. 3.8 Nano-cantilevers in action. A 50-nm-wide cantilever made of a nickel-titanium alloy oscillates after a laser pulse excites it. Blue boxes highlight the movement in 3D. The full movie has one frame every 10 ns. Material properties determined from these oscillations would influence the design of nanomechanical devices. Adapted from Kwon et al. [52]

unprecedented structural detail at a rate that is ten orders of magnitude faster than hitherto (Figs. 3.4, 3.8, and 3.9; see also the review article [24]).

Furthermore, because electrons are focusable and can be pulsed at these very high rates, and because they have appreciable inelastic cross sections, UEM yields information in three distinct ways: in real space, in reciprocal space, and in energy space, all with the changes being followed in the ultrafast time domain. Convergent-beam imaging was shown to provide nanoscale diffraction of heterogeneous ensembles [92], and the power of tomography was also demonstrated for a complex structure [53]. Perhaps the most significant discovery in UEM was the photon induced near-field electron microscopy (PINEM; [3]), which uncovered the nature of the electromagnetic field in nanostructures; shown in Fig. 3.9 are images for PINEM of a single carbon nanotube and the coherent interaction between two particles at nanoscale separations. For biological PINEM imaging [23], see the upcoming Sect. 3.9. Thus, besides structural imaging, the energy landscapes of macromolecules, chemical compositions, and valence and core-energy states can be studied. The 3D structures (from tomography) can also be visualized.

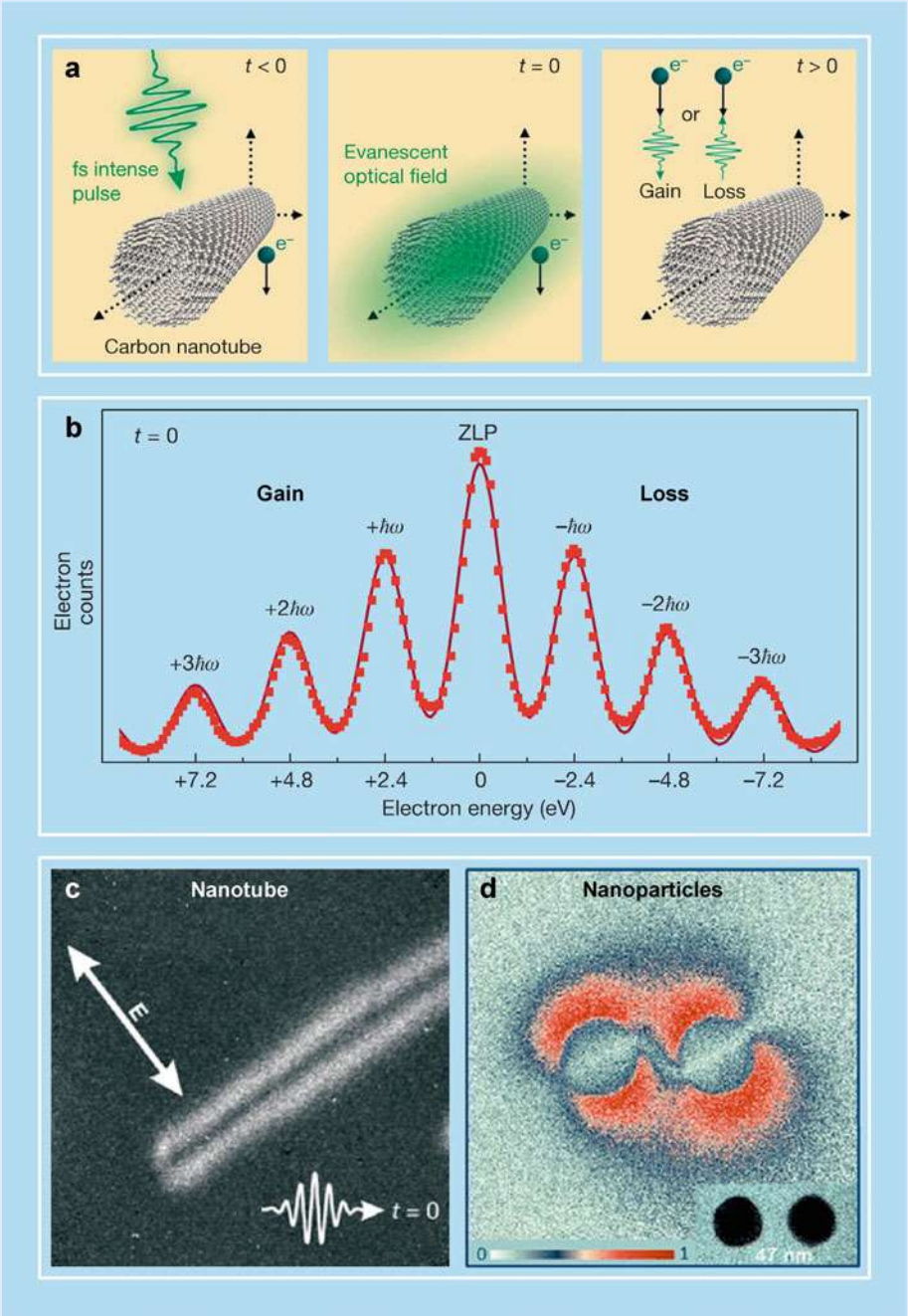


Fig. 3.9 Principles of PINEM and experimental examples. (**a**) *Left frame* shows the moment of arrival of the electron packet at the nanotube prior to the femtosecond laser pulse excitation ($t < 0$); no spatiotemporal overlap has yet occurred. *Middle frame* shows the precise moment at $t = 0$ when the electron packet, femtosecond laser pulse, and evanescent field are at maximum overlap at the carbon nanotube. *Right frame* depicts the process during and immediately after the interaction ($t > 0$) when the electron gains/loses energy equal to integer multiples of femtosecond laser photons. (**b**) PINEM electron energy spectrum obtained at $t = 0$. The spectrum is given in reference to the loss/gain of photon quanta by the electrons with respect to the zero loss peak (ZLP). (**c**) Image taken with the *E*-field polarization of the femtosecond laser pulse being perpendicular to the long-axis of the carbon nanotube at $t = 0$, when the interaction between electrons, photons, and the evanescent field is at a maximum. (**d**) Near-fields of a nanoparticle pair with an edge-to-edge distance of 47 nm with false-color mapping. When the separation between the particles is reduced from 47 to 32 nm, a "channel" is formed between them. Adapted from Barwick et al. [3], Yurtsever et al. [93]

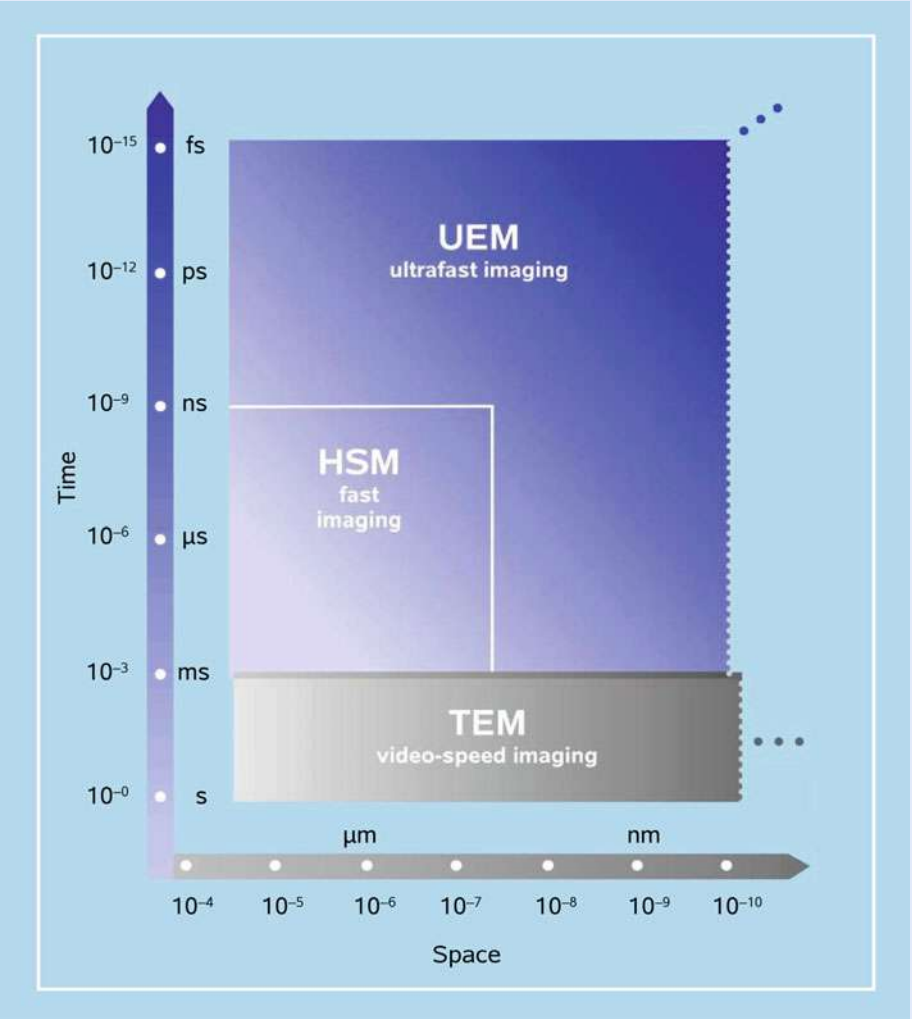


Fig. 3.10 Resolutions in space and time achieved in electron microscopy. The focus here is on the comparison of ultrafast electron microscopy (UEM) and transmission electron microscopy (TEM), but other variants of the techniques (scanning EM, tomography and holography, as well as electron spectroscopy) can similarly be considered. The horizontal dimension represents the spatial resolution achieved from the early years of EM to the era of aberration-corrected instruments. The vertical axis depicts the temporal resolution scale achieved up to the present time and the projected extensions into the near future. The domains of "fast" and "ultrafast" temporal resolutions are indicated by the areas of high-speed microscopy (HSM) and ultrafast electron microscopy (UEM), respectively [98]. Care should be taken in not naming the HSM "ultrafast electron microscopy". Adapted from Zewail [99]

Figure 3.10 depicts the space and time dimensions of UEM, and—for comparison—those of TEM. The boundaries of the time resolution are representative of the transition from the millisecond video speed used in TEM imaging, to the fast, or high-speed nanosecond-to-microsecond imaging, and on to the ultrafast femtosecond-to-attosecond imaging regime. The spatial resolution in the high-speed, nanosecond domain indicated in the Figure is limited by electron–electron (space–charge) repulsion in the nanosecond pulses of electrons. The UEM landscape is that of single-electron imaging, which, owing to the absence of inter-electron repulsion, reaches the spatial resolution of the TEM, but with the temporal resolution being ultrafast. Examples of time-averaged EM and of UEM studies can be found in [98]. The key concepts pertinent to UEM are outlined in Fig. 3.7.

3.8 Coherent Single-Electrons in Ultrafast Electron Microscopy

The concept of single-electron imaging is based on the premise that the trajectories of coherent and timed, single-electron packets can provide an image equivalent to that obtained using many electrons in conventional microscopes. Unlike the random electron distribution of conventional microscopes, in UEM the packets are timed with femtosecond precision, and each electron has a unique coherence volume. As such, each electron of finite de Broglie wavelength is (transversely) coherent over the object length scale to be imaged, with a longitudinal coherence length that depends on its velocity. On the detector, the electron produces a "click" behaving as a classical particle, and when a sufficient number of such clicks are accumulated stroboscopically, the whole image emerges (Fig. 3.7). This was the idea realized in electron microscopy for the first time at Caltech. Putting it in Dirac's famous dictum: *each electron interferes only with itself*. In the microscope, this "stop-motion imaging" yields a real-time movie of the process, and the methodology being used is similar to that described in Sect. 3.3. We note that, in contrast with Muybridge's experiments, which deal with a single object (the horse), here we have to synchronize the motion of numerous independent atoms or molecules so that all of them have reached a similar point in the course of their structural evolution; to achieve such synchronization for millions—or billions—of the studied objects, the relative timing of the clocking and probe pulses must be of femtosecond precision, and the launch configuration must be defined to sub-ångström resolution.

Unlike with photons, in imaging with electrons we must also consider the consequences of the Pauli exclusion principle. The maximum number of electrons that can be packed into a state (or a cell of phase space) is two, one for each spin; in contrast, billions of photons can be condensed in a state of the laser radiation. This characteristic of electrons represents a fundamental difference in what is termed the "degeneracy," or the mean number of electrons per cell in phase space. Typically it is about 10^{-4} to 10^{-6} but it is possible in UEM to increase the degeneracy by orders of magnitude, a feature that could be exploited for studies in quantum electron optics [98]. I note here that the definition of "single-electron packet" is reserved for the case when each timed packet contains one or a small number of electrons such that coulombic repulsion is effectively absent.

At Caltech, three UEM microscopes operate at 30, 120, and 200 keV. Upon the initiation of the structural change by heating of the specimen, or through electronic excitation, using the ultrashort clocking pulse, a series of frames for real space images, and similarly for diffraction patterns or electron-energy-loss spectra (EELS), is obtained. In the single-electron mode of operation, which affords studies of reversible processes or repeatable exposures, the train of strobing electron pulses is used to build up the image. By contrast, in the single-pulse mode, each recorded frame is created with a single pulse that contains 10^4–10^6 electrons. One has the freedom to operate the apparatus in either single-electron or single-pulse mode.

It is known from the Rayleigh criterion that, with a wavelength λ of the probing beam, the smallest distance that can be resolved is given by approximately 0.5 λ. Thus, in conventional optical microscopy, green light cannot resolve distances smaller than approximately 3000 Å (300 nm). Special variants of optical microscopy can nowadays resolve small objects of sub-hundreds of nanometers in size, which is below the diffraction limit ([37] and the references therein). When, however, an electron is accelerated in a 100 kV microscope, its wavelength approaches 4×10^{-2} Å, i.e., the picometer scale, a distance far shorter than that separating the atoms in a solid or molecule. The resolution of an electron

microscope can, in principle, reach the sub-ångström limit [66]. One important advantage in optical microscopy is the ability to study objects with attached chromophores in water. Advances in environmental EM [28] for the studies of catalysis have been achieved, and liquid-cell EM ([17] and the references therein) has been successful in the studies of nanomaterials and cells.

Of the three kinds of primary beams (neutrons, X-rays, and electrons) suitable for structural imaging, the most powerful are coherent electrons, which are readily produced from field-emission guns [98]. The source brightness, as well as the temporal and spatial coherence of such electrons, significantly exceeds the values achievable for neutrons and X-rays: moreover, the minimum probe diameter of an electron beam is as small as 1 Å, and its elastic mean free path is approximately 100 Å (for carbon), much less than for neutrons and X-rays [35]. For larger samples and for those studied in liquids, X-ray absorption spectroscopy, when time-resolved, provides unprecedented details of energy pathways and electronic structural changes [8, 11, 43]. It is significant to note that in large samples the precision is high but it represents an average over the micrometer-scale specimens.

As a result of these developments and inventions, new fields of research continue to emerge. First, by combining energy-filtered electron imaging with electron tomography, chemical compositions of sub-attogram (less than 10^{-18} g) quantities located at the interior of microscopic or mesoscopic objects may be retrieved non-destructively. Second, transmission electron microscopes fitted with field-emission guns to provide coherent electron rays can be readily adapted for electron holography to record the magnetic fields within and surrounding nanoparticles or metal clusters, thereby yielding the lines of force of, for example, a nanoferromagnet encapsulated within a multi-walled carbon nanotube. Third, advances in the design of aberration-corrected high-resolution EMs have greatly enhanced the quality of structural information pertaining to nanoparticle metals, binary semiconductors, ceramics, and complex oxides. Moreover, electron tomography sheds light on the shape, size, and composition of materials. Finally, with convergent-beam and near-field 4D UEM [3, 92], the structural dynamics and plasmonics of a nanoscale single site (particle), and of nanoscale interface fields, can be visualized, reaching unprecedented resolutions in space and time [98].

3.9 Visualization and Complexity

Realization of the importance of visualization and observation is evident in the exploration of natural phenomena, from the very small to the very large. A century ago, the atom appeared complex, a "raisin or plum pie of no structure," until it was visualized on the appropriate length and time scales. Similarly, with telescopic observations, a central dogma of the cosmos was changed and complexity yielded to the simplicity of the heliocentric structure and motion in the entire Solar System. From the atom to the universe, the length and time scales span extremes of powers of 10. The electron in the first orbital of a hydrogen atom has a "period" of sub-femtoseconds, and the size of atoms is on the nanometer scale or less. The lifetime of our universe is approximately 13 billion years and, considering the light year (approx. 10^{16} m), its length scale is of the order of 10^{26} m. In between these scales lies the world of life processes, with scales varying from nanometers to centimeters and from femtoseconds to seconds.

In the early days of DNA structural determination (1950s), a cardinal concept, in vogue at that time, was encapsulated in Francis Crick's statement: *If you want to know the function, determine the structure.* This view dominated the thinking at the time, and it was what drove Max Perutz and John Kendrew earlier in their studies of proteins. But as we learn more about complexity, it becomes clear that

the so-called structure–function correlation is insufficient to establish the mechanisms that determine the behavior of complex systems [98]. For example, the structures of many proteins have been determined, but we still do not understand how they fold, how they selectively recognize other molecules, how the matrix water assists folding and the role it plays in directionality, selectivity, and recognition; see, e.g., [61] for protein behavior in water (hydrophobic effect) and [62] for complexity, even in isolated systems. The proteins hemoglobin and myoglobin (a subunit of hemoglobin) have unique functions: the former is responsible for transporting oxygen in the blood of vertebrates, while the latter carries and stores oxygen in muscle cells. The three-dimensional structures of the two proteins have been determined (by Perutz and Kendrew), but we still do not understand the differences in behavior in the oxygen uptake by these two related proteins, the role of hydration, and the exact nature of the forces that control the dynamics of oxygen binding and liberation from the haem group. Visualization of the changing structures during the course of their functional operation is what is needed (see, e.g., [12, 75]).

In biological transformations, the energy landscape involves very complicated pathways, including those that lead to a multitude of conformations, with some that are "active" and others that are "inactive" in the biological function. Moreover, the landscapes define "good" and "bad" regions, the latter being descriptive of the origin of molecular diseases. It is remarkable that the robustness and function of these "molecules of life" are the result of a balance of weak forces—hydrogen bonding, electrostatic forces, dispersion, and hydrophobic interactions—all of energy of the order of a few kcal·mol^{-1}, or approximately 0.1 eV or less. Determination of time-averaged molecular structures is important and has led to an impressive list of achievements, for which more than ten Nobel Prizes have been awarded, but the structures relevant to function are those that exist in the non-equilibrium state. Understanding their behavior requires an integration of the trilogy: structure, dynamics, and function.

Figure 3.11 depicts the experimental PINEM field of *Escherichia coli* bacterium which decays on the femtosecond time scale; in the same Figure, we display the conceptual framework of cryo-UEM for the study of folding/unfolding in proteins. Time-resolved cryo-EM has been successfully introduced in the studies of amyloids [21, 22]. In parallel, theoretical efforts ([57–60, 98] and the references therein) have been launched at Caltech to explore the areas of research pertaining to biological structures, dynamics, and the energy landscapes, with focus on the elementary processes involved.

Large-scale complexity is also evident in correlated physical systems exhibiting, e.g., superconductivity, phase transitions, or self-assembly, and in biological systems with emergent behavior [96]. For materials, an assembly of atoms in a lattice can undergo a change, which leads to a new structure with properties different from the original ones. In other materials, the structural transformation leads to a whole new material phase, as in the case of metal–insulator phase transitions. Questions of fundamental importance pertain to the time and length scales involved and to the elementary pathways that describe the mechanism. Recently, a number of such questions have been addressed by means of 4D electron imaging. Of significance are two regimes of structural transformation: the first one involves an initial (coherent) bond dilation that triggers unit-cell expansion and phase growth [5, 9], and the second one involves phase transformations in a diffusionless (collective) process that emerges from an initial random motion of atoms [67]. The elementary processes taking place in superconducting materials are now being examined [26] by direct probing of the "ultrafast phonons," which are critically involved, and—in this laboratory—by studying the effect of optical excitation [10, 29] in ultrafast electron crystallography (UEC).

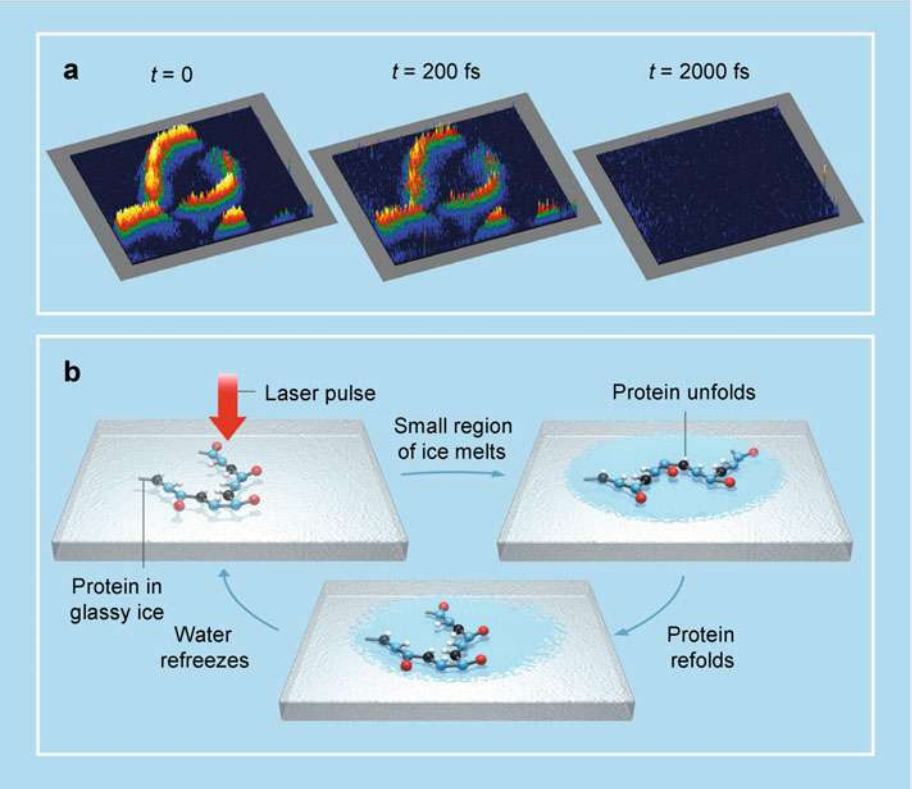


Fig. 3.11 Biological UEM. (**a**) An *Escherichia coli* bacterium imaged with PINEM. A femtosecond laser pulse generates an evanescent electromagnetic field in the cell's membrane at time zero. By collecting only the imaging electrons that gained energy from this field, the technique produces high-contrast, relatively high-spatial resolution snapshot of the membrane. The false-color contour plot depicts the intensity recorded. The method can capture events occurring on very short timescales, as is evinced by the field's significant decay after 200 fs. The field vanishes by 2000 fs. (**b**) By adapting the technique called cryo-imaging, we proposed the use of 4D UEM for the observation of biological processes such as protein folding. A glassy (noncrystalline) ice holds the protein. For each shot of the movie, a laser pulse melts the ice around the sample, causing the protein to unfold in the warm water. The movie records the protein refolding before the water cools and refreezes. The protein could be anchored to the substrate to keep it in the same position for each shot. Adapted from Flannigan et al. [23], Zewail [101]

3.10 Attosecond Pulse Generation

For photons, attosecond pulse generation has already been demonstrated [49], and with a unique model that describes the processes involved [13]. Several review articles have detailed the history shaped by many involved in the field and the potential for applications; I recommend the reviews by Corkum [44], Krausz and Stockman [50], and Vrakking [55], and the critique by Leone et al. [54]. It is true that the pulse width can approach the 100 attosecond duration or less (see Fig. 3.12 for the electric field pattern), yielding a few-cycles-long pulse. However, there is a price to pay; the band width in the energy domain becomes a challenge in designing experiments. A 20-attosecond pulse has an associated energy bandwidth of $\Delta E \approx 30$ eV. In the femtochemistry domain, the bandwidth permits the mapping of dynamics on a given potential energy surface, and with selectivity for atomic motions. With electrons, processes of ionization and electron density change can be examined, but not with the selectivity mentioned above, as the pulse energy width covers, in case of molecules, numerous energy states. Experimental success has so far been reported for ionized atoms, Auger processes, and the direct measurement of the current produced by valence–conduction-band excitation in SiO_2.

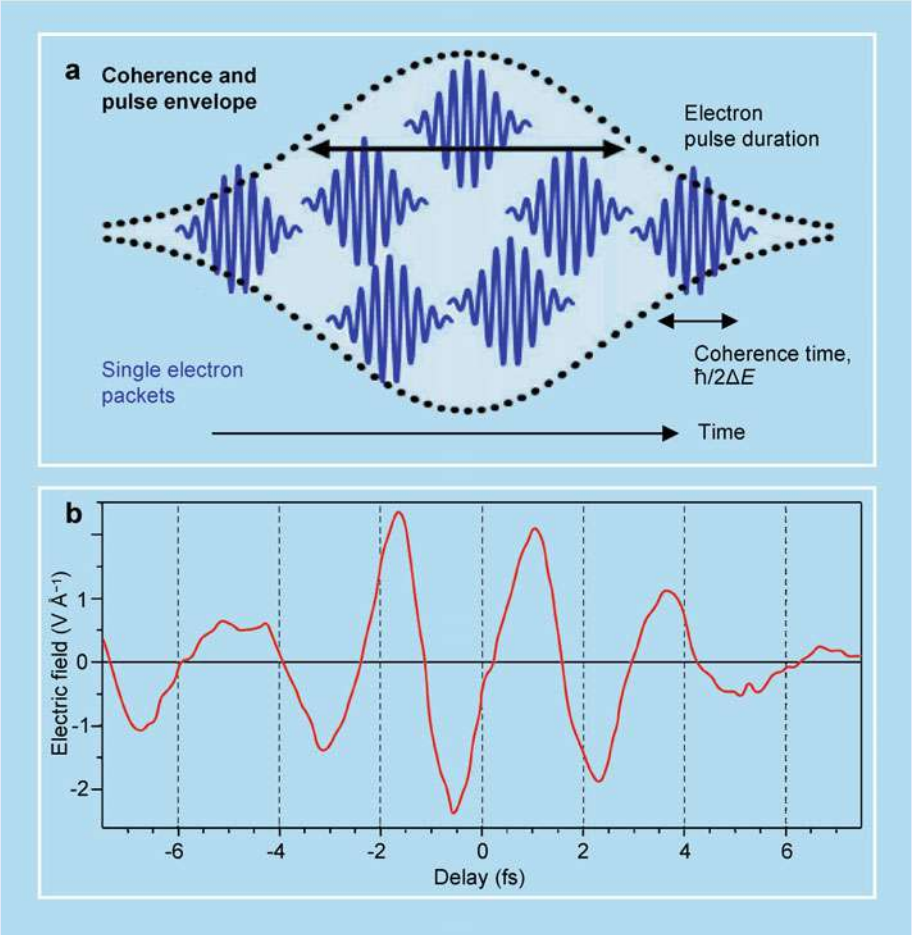


Fig. 3.12 Electron pulse coherence and its packets, together with measured attosecond pulse envelope. (**a**) Single-electron packets and electron pulses. Shown are the effective pulse parameters and the coherence time involved. Each single-electron (*blue*) is a coherent packet consisting of many cycles of the de Broglie wave and has different timing due to the statistics of generation. On average, multiple single-electron packets form an effective electron pulse (dotted envelope). (**b**) Electric field of a few-cycle laser pulse impinging on an SiO_2 sample. Adapted from Baum and Zewail [7], Krausz and Stockman [50]

On the other hand, for electrons in UEM, the challenge is to push the limit of the temporal width into the femtosecond domain and on to the attosecond regime. Several schemes have been proposed and discussed in the review article by Baum and Zewail [7]. In Fig. 3.12 we display a schematic for a single-electron packet and pulse. The electron trajectories obtained for the temporal optical-grating, tilted-pulses, and temporal-lens methodology are given in Fig. 3.13. One well-known technique is that of microwave compression [20] of electron pulses, which has recently been applied in femtosecond electron diffraction setups [63].

A more promising method for compression directly to the attosecond domain involves the creation of "temporal lenses" made by ultrashort laser pulses [6, 38]. The technique relies on the ponderomotive force (or ponderomotive potential) that influences electrons when they encounter an intense electromagnetic field. To create trains of attosecond electron pulses, appropriate optical intensity patterns have to be synchronized with the electron pulse. This is done by using counter-propagating laser pulses to create a standing optical wave that must be both spatially and temporally overlapped with the femtosecond electron pulse to get the desired compression. To make the standing wave in the rest frame

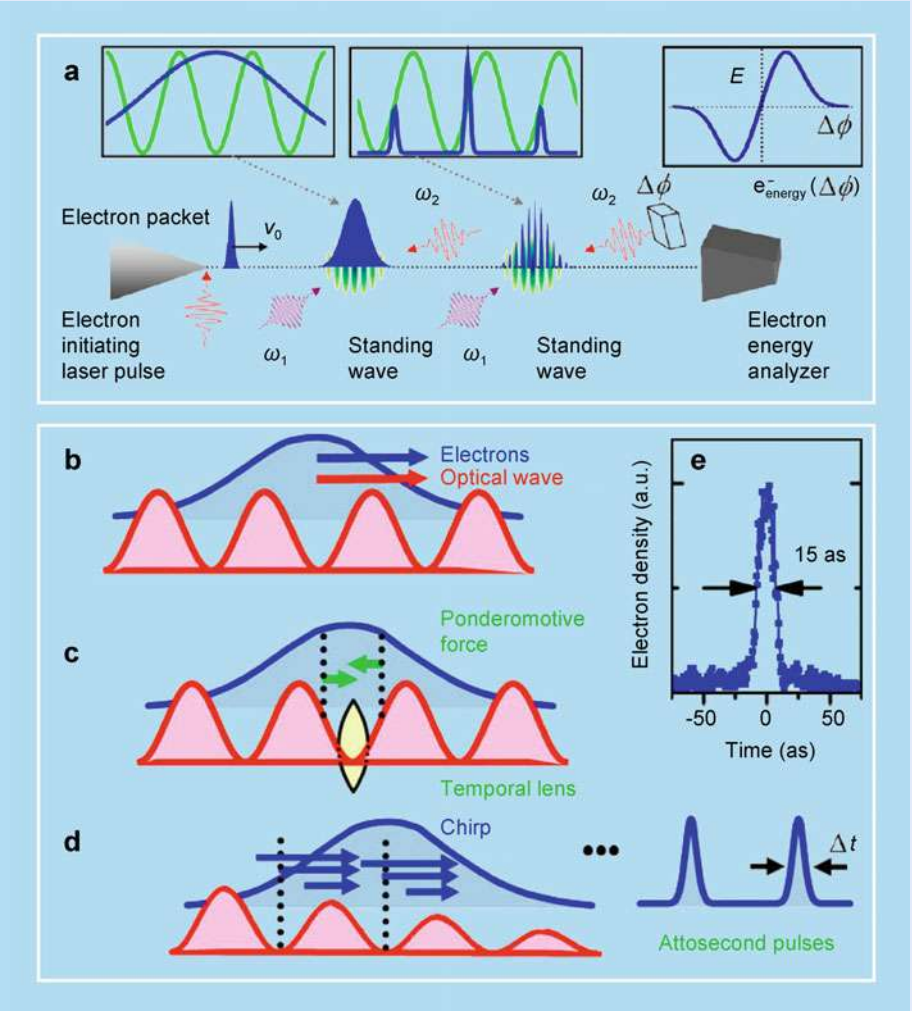


Fig. 3.13 Schemes for creating and measuring attosecond compressed electron packets. (**a**) Temporal lensing. To measure the duration of the attosecond pulses, a second co-propagating standing wave is made to coincide with the electron pulse at the focal position. Instead of using a temporal delay, a phase shift, $\Delta\varphi$, is introduced into one of the laser pulses that creates the probing standing wave. By varying this phase shift, the nodes of the standing wave shift position. The average electron energy can thus be plotted vs. this phase shift. As the electron pulses become shorter than the period of the standing wave, the change in the average energy will increase. To use the attosecond electron pulse train as a probing beam in a UEM, the specimen would need to be positioned where the second standing wave appears in the figure; see Sec. 3.10. (**b**–**e**) Temporal optical gratings for the generation of free attosecond electron pulses for use in diffraction. (**b**) A femtosecond electron packet (*blue*) is made to co-propagate with a moving optical intensity grating (*red*). (**c**) The ponderomotive force pushes electrons toward the minima and thus creates a temporal lens. (**d**) The induced electron chirp leads to compression to the attosecond duration at a later time. (**e**) The electron pulse duration from 10^5 trajectories reaches into the domain of few attoseconds. Adapted from Barwick and Zewail [4], Baum and Zewail [7]

of the electron pulse, the two counter-propagating electromagnetic waves must have different frequencies [6] or be angled appropriately [38]; see Fig. 3.13.

The standing wave that appears in the rest frame of the traveling electron pulse introduces a series of high and low intensity regions, and in this periodic potential each of the individual ponderomotive potential "wells" causes a compression of the local portion of the electron pulse that encounters it [38]. After interaction with the optical potential well, the electrons that have encountered steep intensity gradients get sped up or slowed down, depending on their position in the potential. After additional propagation, the electron pulse self-compresses into a train of attosecond pulses, with the pulse train spacing being equal to the periodicity of the optical standing wave. By placing the compression potential at an appropriate

distance before the specimen, the pulse maximally compresses when encountering the system under study.

3.11 Optical Gating of Electrons and Attosecond Electron Microscopy

As discussed in this chapter, PINEM can be used to generate attosecond pulses [7]. Theoretically, it was clear that spatial and temporal coherences of partial waves (the so-called Bessel functions) can produce attosecond pulse trains, whereas incoherent waves cannot [68–70]. With this in mind, such pulse trains have been generated [19, 47], and the coherent interference of a plasmonic near-field was visualized [71]. Very recently at Caltech, we have developed a new variant of PINEM, which constitutes a breakthrough in electron pulse slicing and imaging.

In all the previous experiments conducted in 4D ultrafast electron microscopy, only "*one optical pulse*" is used to initiate the change in the nanostructure. In a recent report [32], based on the conceptual framework given by Park and Zewail [70], we have used "*two optical pulses*" for the excitation and "*one electron pulse*" for probing. The implementation of this pulse sequence led to the concept of "photon gating" of electron pulses, as shown in Fig. 3.14. The sequence gives rise

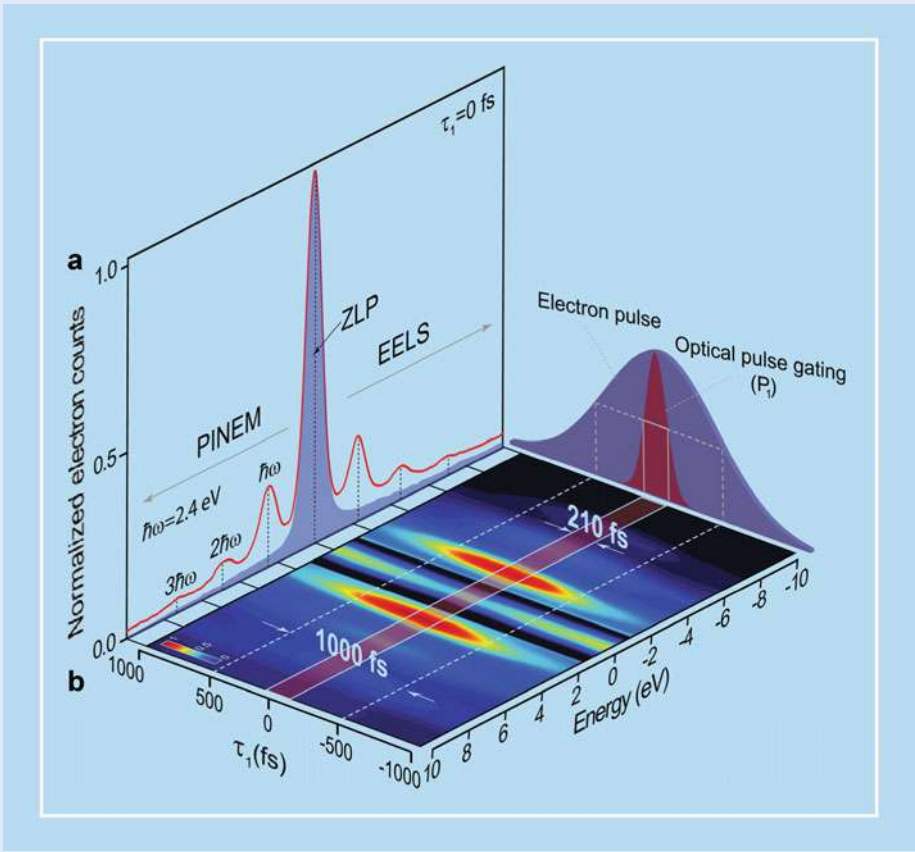


Fig. 3.14 Ultrafast "optical gating" of electrons using three-pulse sequence in PINEM. (**a**) PINEM spectrum at $\tau_1 = 0$ fs, which consists of discrete peaks on the higher and lower energy sides of the zero loss peak (ZLP) separated by multiple photon-energy quanta (2.4 eV). The shaded curve presents the normalized ZLP measured at $\tau_1 = 1000$ fs. (**b**) PINEM spectrogram of photon–electron coupling of the first optical and electron pulse as a function of the first optical pulse delay (τ_1). The ZLP area between −1.5 eV and 1.5 eV has been reduced for visualization of the adjacent discrete peaks. Optical gating is clearly manifested in the narrow strip corresponding to the width of the optical pulse (210 ± 35 fs) shown in red in the vertical plane at right, which is superimposed on the ultrafast electron pulse (1000 fs) in blue. The material studied is vanadium dioxide nanoparticles which undergo metal-to-insulator phase transition when appropriately excited. Adapted from Hassan et al. [32]

to an electron pulse width limited only by the optical-gate pulse width. A picosecond electron pulse was shown to compress into the femtosecond width of the exciting optical pulse. This is a very important advance with the potential for reaching the attosecond time domain with many applications in 4D materials visualization.

3.12 Conclusion

The generation of femtosecond light has its origins in several advances: mode locking by Anthony DeMaria, dye lasers by Peter Sorokin and Fritz Schäfer, colliding pulses in dye lasers by Erich Ippen and Chuck Shank (see the books edited by Schäfer and Shapiro [74, 76]); chirped pulse amplification by Gérard Mourou; white-light continuum by Bob Alfano and colleagues; and the streaking methods by Dan Bradley, and others, have all contributed to the advances made. Prior to the development of femtochemistry, there were significant contributions made in picosecond chemistry and physics (by Peter Rentzepis, Ken Eisenthal, the late Robin Hochstrasser, and Wolfgang Kaiser). In our laboratories at Caltech, the development of *femtochemistry* in the 1980s involved the use of spectroscopy, photoelectron detection, and mass spectrometry. But nothing can be compared with the capability we developed from the year 2000 and until now for microscopy imaging in 4D UEM.

The microscope is arguably one of the two most powerful human-made instruments of all time, the other being the telescope. To our vision they brought the very small and the very distant. Robert Hooke, for his *Micrographia*, chose the subtitle: *or some physiological descriptions of minute bodies made by magnifying glasses with observations and inquiries thereupon*. These words were made in reference to conventional optical microscopes; the spatial resolution of them is being limited by the wavelength of visible light—the Rayleigh criterion, as mentioned above. The transmission electron microscope, since its invention in the 1930s, has provided the wavelength of picometers, taking the field of imaging beyond the "minutes" of the seventeenth century *Micrographia*—it has now become possible to image individual atoms, and the scope of applications spans essentially all of the physical sciences as well as biology.

With 4D ultrafast electron microscopy, the structures determined are no longer "time-averaged" over seconds of recording. They can be seen as frames of a movie that elucidates the nature of the processes involved. We have come a long way from the epochs of the *camera obscura* of Alhazen and Hooke's *Micrographia*, but I am confident that new research fields will continue to emerge in the twenty-first century, especially within frontiers at the intersection of physical, chemical, and biological sciences [97]. Indeed, the microscopic invisible has become visible—thanks to ultrafast photons and electrons.

Acknowledgements The research summarized in this contribution had been carried out with support from the National Science Foundation (DMR-0964886) and the Air Force Office of Scientific Research (FA9550-11-1-0055) in the Physical Biology Center for Ultrafast Science and Technology (UST), which is supported by the Gordon and Betty Moore Foundation at Caltech.

During the forty-years-long research endeavor at Caltech, I had the pleasure of working with some 400 research associates, and without their efforts the above story would not have been told. References are included here to highlight selected contributions, but the work in its totality could not be covered because of the limited space and the article focus.

I am especially grateful to Dr. Dmitry Shorokhov, not only for his technical support but also for the intellectual discussions of the science involved and possible future projects. Dr. Dmitry Shorokhov, together with Dr. Milo Lin, has made major contributions to biological dynamics in the isolated phase (see Sec. 3.9).

References

1. Al-Hassani STS, Woodcock E, Saoud R (eds) (2006) 1001 inventions: Muslim heritage in our world. Foundation for Science, Technology and Civilization, Manchester
2. Amunts A, Brown A, Bai X et al (2014) Structure of the yeast mitochondrial large ribosomal subunit. Science 343:1485–1489
3. Barwick B, Flannigan DJ, Zewail AH (2009) Photon induced near-field electron microscopy. Nature 462:902–906
4. Barwick B, Zewail AH (2015) Photonics and plasmonics in 4D ultrafast electron microscopy. ACS Photonics 2:1391–1402
5. Baum P, Yang DS, Zewail AH (2007) 4D visualization of transitional structures in phase transformations by electron diffraction. Science 318:788–792
6. Baum P, Zewail AH (2007) Attosecond electron pulses for 4D diffraction and microscopy. Proc Natl Acad Sci U S A 104:18409–18414
7. Baum P, Zewail AH (2009) 4D attosecond imaging with free electrons: diffraction methods and potential applications. Chem Phys 366:2–8
8. Bressler C, Milne C, Pham VT et al (2009) Femtosecond XANES study of the light-induced spin crossover dynamics in an iron(II) complex. Science 323:489–492
9. Cavalleri A (2007) All at once. Science 318:755–756
10. Carbone F, Yang DS, Giannini E et al (2008) Direct role of structural dynamics in electron-lattice coupling of superconducting cuprates. Proc Natl Acad Sci USA 105:20161–20166
11. Chergui M, Zewail AH (2009) Electron and X-ray methods of ultrafast structural dynamics: advances and applications. Chem Phys Chem 10:28–43
12. Cho HS, Schotte F, Dashdorj N et al (2013) Probing anisotropic structure changes in proteins with picosecond time-resolved small-angle X-ray scattering. J Phys Chem B 117:15825–15832
13. Corkum PB, Krausz F (2007) Attosecond science. Nat Phys 3:381–387
14. Cowley JM (1995) Diffraction physics, 3rd edn. Elsevier, Amsterdam
15. Crewe AV, Wall J, Langmore J (1970) Visibility of single atoms. Science 168:1338–1340
16. Crowther RA (2008) The Leeuwenhoek lecture 2006. Microscopy goes cold: frozen viruses reveal their structural secrets. Philos Trans R Soc B 363:2441–2451
17. de Jonge N, Peckys DB, Kremers GJ et al (2009) Electron microscopy of whole cells in liquid with nanometer resolution. Proc Natl Acad Sci U S A 106:2159–2164
18. Dantus M, Kim SB, Williamson JC et al (1994) Ultrafast electron diffraction V: experimental time resolution and applications. J Phys Chem 98:2782–2796
19. Feist A, Echternkamp KE, Schauss J et al (2015) Quantum coherent optical phase modulation in an ultrafast transmission electron microscope. Nature 521:200–203

20. Fill E, Veisz L, Apolonski A et al (2006) Sub-fs electron pulses for ultrafast electron diffraction. New J Phys 8:272
21. Fitzpatrick AWP, Lorenz UJ, Vanacore GM et al (2013) 4D Cryo-electron microscopy of proteins. J Am Chem Soc 135:19123–19126
22. Fitzpatrick AWP, Park ST, Zewail AH et al (2013) Exceptional rigidity and biomechanics of amyloid revealed by 4D electron microscopy. Proc Natl Acad Sci USA 110:10976–10981
23. Flannigan DJ, Barwick B, Zewail AH (2010) Biological imaging with 4D ultrafast electron microscopy. Proc Natl Acad Sci U S A 107:9933–9937
24. Flannigan DJ, Zewail AH (2012) 4D electron microscopy: principles and applications. Acc Chem Res 45:1828–1839
25. Forsén S (1992) The Nobel prize for chemistry. In: Frängsmyr T, Malmstrøm BG (eds) Nobel lectures in chemistry, 1981–1990. World Scientific, Singapore, p 257, transcript of the presentation made during the 1986 Nobel Prize in chemistry award ceremony
26. Först M, Mankowsky R, Cavalleri A (2015) Mode-selective control of the crystal lattice. Acc Chem Res 48:380–387
27. Frank J (2006) Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. Oxford University Press, New York
28. Gai PL, Boyes ED (2003) Electron microscopy in heterogeneous catalysis. Series in microscopy in materials science. IOP Publishing, Bristol
29. Gedik N, Yang DS, Logvenov G et al (2007) Nonequilibrium phase transitions in cuprates observed by ultrafast electron crystallography. Science 316:425–429
30. Glaeser RM (2008) Macromolecular structures without crystals. Proc Natl Acad Sci U S A 105:1779–1780
31. Glaeser RM, Downing K, DeRosier D et al (2007) Electron crystallography of biological macromolecules. Oxford University Press, New York
32. Hassan MT, Liu H, Baskin JS et al (2015) Photon gating in four-dimensional ultrafast electron microscopy. Proc Natl Acad Sci U S A 112:12944–12949
33. Hawkes PW (2009) Aberration correction: past and present. Phil Trans R Soc A 367:3637–3664
34. Hawkes PW, Spence JCH (eds) (2007) Science of microscopy. Springer, New York
35. Henderson R (1995) The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. Q Rev Biophys 28:171–193
36. Henderson R (2002) Excitement over X-ray lasers is excessive. Nature 415:833
37. Hell SW (2015) Nanoscopy with focused light. Angew Chem Int Ed 54:8054–8066, transcript of the Nobel lecture given in 2014
38. Hilbert SA, Uiterwaal C, Barwick B et al (2009) Temporal lenses for attosecond and femtosecond electron pulses. Proc Natl Acad Sci U S A 106:10558–10563
39. Hooke R (1665) Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses with observations and inquiries thereupon. Royal Society, London
40. Howie A (2009) Aberration correction: zooming out to overview. Phil Trans R Soc A 367:3859–3870
41. Humphreys CJ (ed) (2002) Understanding materials: a festschrift for Sir Peter Hirsch. Maney, London
42. Jain VK (1995) The world's fastest camera. The World and I 10:156–163
43. Kim TK, Lee JH, Wulff M et al (2009) Spatiotemporal kinetics in solution studied by time-resolved X-ray liquidography (solution scattering). Chem Phys Chem 10:1958–1980
44. Kim KT, Villeneuve DM, Corkum PB (2014) Manipulating quantum paths for novel attosecond measurement methods. Nat Photonics 8:187–194
45. Klug A (1983) From macromolecules to biological assemblies. Angew Chem Int Ed 22:565–636, transcript of the Nobel lecture given in 1982
46. Knoll M, Ruska E (1932) Das Elektronenmikroskop. Z Phys 78:318–339
47. Kociak M (2015) Microscopy: quantum control of free electrons. Nature 521:166–167
48. Kornberg R (2007) The molecular basis of eukaryotic transcription. Angew Chem Int Ed 46:6956–6965, transcript of the Nobel lecture given in 2006
49. Krausz F, Ivanov M (2009) Attosecond physics. Rev Mod Phys 81:163–234
50. Krausz F, Stockman MI (2014) Attosecond metrology: from electron capture to future signal processing. Nat Photonics 8:205–213
51. Kühlbrandt W (2014) The resolution revolution. Science 343:1443–1444
52. Kwon OH, Park HS, Baskin JS et al (2010) Nonchaotic, nonlinear motion visualized in complex nanostructures by stereographic 4D electron microscopy. Nano Lett 10:3190–3198
53. Kwon OH, Zewail AH (2010) 4D electron tomography. Science 328:1668–1673

54. Leone SR, McCurdy CW, Burgdörfer J et al (2014) What will it take to observe processes in 'real time'? Nat Photonics 8:162–166
55. Lépine F, Ivanov MY, Vrakking MJJ (2014) Attosecond molecular dynamics: fact or fiction? Nat Photonics 8:195–204
56. Lichte H (2002) Electron interference: mystery and reality. Philos Trans R Soc Lond A 360:897–920
57. Lin MM, Shorokhov D, Zewail AH (2006) Helix-to-coil transitions in proteins: helicity resonance in ultrafast electron diffraction. Chem Phys Lett 420:1–7
58. Lin MM, Meinhold L, Shorokhov D et al (2008) Unfolding and melting of DNA (RNA) hairpins: the concept of structure-specific 2D dynamic landscapes. Phys Chem Chem Phys 10:4227–4239
59. Lin MM, Shorokhov D, Zewail AH (2009) Structural ultrafast dynamics of macromolecules: diffraction of free DNA and effect of hydration. Phys Chem Chem Phys 11:10619–10632
60. Lin MM, Shorokhov D, Zewail AH (2009) Conformations and coherences in structure determination by ultrafast electron diffraction. J Phys Chem A 113:4075–4093
61. Lin MM, Zewail AH (2012) Protein folding: simplicity in complexity. Ann Phys 524:379–391
62. Lin MM, Shorokhov D, Zewail AH (2014) Dominance of misfolded intermediates in the dynamics of α-helix folding. Proc Natl Acad Sci U S A 111:14424–14429
63. Mancini GF, Mansart B, Pagano S et al (2012) Design and implementation of a flexible beamline for fs electron diffraction experiments. Nucl Instrum Methods Phys Res A 691:113–122
64. Müller EW (1951) Das Feldionenmikroskop. Z Phys 131:136–142
65. Nellist PD, Chisholm MF, Dellby N et al (2004) Direct sub-ångström imaging of a crystal lattice. Science 305:1741
66. O'Keefe MA (2008) Seeing atoms with aberration-corrected sub-ångström electron microscopy. Ultramicroscopy 108:196–209
67. Park HS, Kwon OH, Baskin JS et al (2009) Direct observation of martensitic phase-transformation dynamics in iron by 4D single-pulse electron microscopy. Nano Lett 9:3954–3962
68. Park ST, Lin MM, Zewail AH (2010) Photon induced near-field electron microscopy (PINEM): theoretical and experimental. New J Phys 12:123028
69. Park ST, Kwon OH, Zewail AH (2012) Chirped imaging pulses in four-dimensional electron microscopy: femtosecond pulsed hole burning. New J Phys 14:053046
70. Park ST, Zewail AH (2012) Enhancing image contrast and slicing electron pulses in 4D near-field electron microscopy. Chem Phys Lett 521:1–6
71. Piazza L, Lummen TTA, Quiñonez E et al (2015) Simultaneous observation of the quantization and the interference pattern of a plasmonic near-field. Nat Commun 6:6407
72. Sabra AI (2003) Ibn al-Haytham. Harvard Mag 9–10:54–55
73. Sali A, Glaeser RM, Earnest T et al (2003) From words to literature in structural proteomics. Nature 422:216–225
74. Schäfer FP (ed) (1973) Dye lasers (Topics in applied physics, vol. 1) Springer, Heidelberg
75. Schotte F, Cho HS, Kaila VRI et al (2012) Watching a signaling protein function in real time via 100-ps time-resolved Laue crystallography. Proc Natl Acad Sci USA 109:19256–19261
76. Shapiro SL (ed) (1977) Ultrashort light pulses, picosecond techniques and applications (Topics in applied physics, vol. 18) Springer, Heidelberg
77. Silverman MP, Strange W, Spence JCH (1995) The brightest beam in science: new directions in electron microscopy and interferometry. Am J Phys 63:800–813
78. Spence JCH (2003) High-resolution electron microscopy, (Monographs on the physics and chemistry of materials, vol. 60), 3rd edn. Oxford University Press, New York
79. Spence JCH (2009) Electron interferometry. In: Greenberger D, Hentschel K, Weinert F (eds) Compendium of quantum physics: concepts, experiments, history and philosophy. Springer, Berlin, pp 188–195
80. Steffens B (2007) Ibn al-Haytham: first scientist. Morgan Reynolds, Greensboro
81. Strutt JW (1891) On pin-hole photography. Philos Mag 31:87–99
82. Thomas JM (1979) Direct imaging of atoms. Nature 281:523–524
83. Thomas JM (1991) Femtosecond diffraction. Nature 351:694–695
84. Thomas JM (2004) Ultrafast electron crystallography: the dawn of a new era. Angew Chem Int Ed 43:2606–2610
85. Thomas JM (2005) A revolution in electron microscopy. Angew Chem Int Ed 44:5563–5566

86. Thomas JM (2008) Revolutionary developments from atomic to extended structural imaging. In: Zewail AH (ed) Physical biology: from atoms to medicine. Imperial College Press, London, pp 51–114
87. Thomas JM (2009) The renaissance and promise of electron energy-loss spectroscopy. Angew Chem Int Ed 48:8824–8826
88. Tonomura A (1998) The quantum world unveiled by electron waves. World Scientific, Singapore
89. Tonomura A (1999) Electron holography, 2nd edn. Springer, Berlin
90. Tsong TT (2006) Fifty years of seeing atoms. Phys Today 59:31–37
91. Williamson JC, Cao J, Ihee H et al (1997) Clocking transient chemical changes by ultrafast electron diffraction. Nature 386:159–162
92. Yurtsever A, Zewail AH (2009) 4D nanoscale diffraction observed by convergent-beam ultrafast electron microscopy. Science 326:708–712
93. Yurtsever A, Baskin JS, Zewail AH (2012) Entangled nanoparticles: discovery by visualization in 4D electron microscopy. Nano Lett 12:5027–5032
94. Zewail AH (1990) The birth of molecules. Sci Am 263:76–82
95. Zewail AH (2000) Femtochemistry: atomic-scale dynamics of the chemical bond using ultrafast lasers. In: Frängsmyr T (ed) Les prix nobel: the nobel prizes 1999. Almqvist & Wiksell, Stockholm, pp 110–203, Also published in Angew Chem Int Ed 39:2587–2631; transcript of the Nobel lecture given in 1999
96. Zewail AH (2008) Physical biology: 4D visualization of complexity. In: Zewail AH (ed) Physical biology: from atoms to medicine. Imperial College Press, London, pp 23–49
97. Zewail AH (2009) Chemistry at a historic crossroads. Chem Phys Chem 10:23
98. Zewail AH, Thomas JM (2009) 4D electron microscopy: imaging in space and time. Imperial College Press, London
99. Zewail AH (2010) 4D electron microscopy. Science 328:187–193
100. Zewail AH (2010) Micrographia of the 21st century: from camera obscura to 4D microscopy. Philos Trans R Soc Lond A 368:1191–1204
101. Zewail AH (2010) Filming the invisible in 4D. Sci Am 303:74–81
102. Zewail AH (2012) 4D imaging in an ultrafast electron microscope. US Patent 8,203,120
103. Zewail AH (2014) 4D visualization of matter: recent collected works. Imperial College Press, London

Optical Sources

Contents

The Laser

Bahaa Saleh

B. Saleh (✉)
CREOL, The College of Optics and Photonics, University of Central Florida, Orlando, FL, USA
e-mail: besaleh@creol.ucf.edu

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_4

4

One of the greatest inventions of the twentieth century, if not of all times, the laser is regarded as humankind's most versatile light source, a truly new kind of light with remarkable properties unlike anything that existed before—a light fantastic! Since it was first used to generate light in 1960 by Theodore Maiman at Hughes Research Laboratories with the help of a ruby crystal, the laser has been at the core of most light-based technologies and has garnered applications in all facets of life. It is a marvelous tool that has enabled many scientific discoveries. Numerous books, textbooks, and articles have been written about laser physics and engineering. This article is a brief tutorial introducing some of the basic principles underlying the development of the laser and highlighting some of its remarkable characteristics.

4.1 Introduction: A Laser in the Hands of Ibn al-Haytham

As the millennium-old contributions of al-Hasan ibn al-Haytham (ca. 965–ca. 1040) to light and vision are being celebrated in this *International Year of Light*, one might wonder if the laser could have been invented in al-Hasan's time. It is even tempting to imagine him being handed a red light beam from a laser, shining through the smoke created by incense somewhere in Fatimid Cairo, and to speculate on his possible reaction. As the reigning expert of optics and vision of his time, would he have found laser light to be truly remarkable? Would he have used it to corroborate his original observations pertaining to reflection, refraction, and the focusing of light?

Seeing the laser beam, al-Hasan would probably have surmised that it could be sunlight shaped into a thin beam by the use of some ingenious contraption of mirrors, a tiny version of the mirror systems said to have been used by Archimedes to destroy the Roman fleet in 212 BC. What about the red color of the laser beam? Simple: it's sunlight transmitted through a piece of red glass, much like those made and traded by the Phoenician glassmakers. Puzzled about the unusual thinness of the beam, its very limited divergence, and its exceptional brightness, al-Hasan would have waited for the Sun to set; seeing that the light beam still shined brilliantly through the smoke, he might then have concluded that the contraption uses a miniaturized version of a red lantern, similar to those he saw as a child in Basra. He would then have again contemplated the exceptional brightness of the light. And as to the narrowness of the beam, he might have speculated that it is collimated light passed through tiny holes much like those used in the then-known camera obscura.

al-Hasan Ibn al-Haytham
(Alhazen) ca. 965 – ca. 1040

Known for his methodical reliance on experimentation and controlled testing, al-Hasan would have conducted an experiment using the brightest red lantern available, along with mirrors, magnifying lenses, and pinholes, to produce a similar

beam of light. He probably would have succeeded in creating a dim red beam of light, a miniaturized version of the beam produced at the Pharos (lighthouse) of Alexandria, which used a mirror to reflect sunlight during the day and a fire lit at night. But no matter what he might have done with the oil-based light sources of the day, it would not have been possible for him to come close to the brightness and narrowness of the laser beam.

Frustrated with his failure, al-Hasan would probably have engaged in a set of experiments benefitting from the available "magic" light source to confirm his findings about reflection and refraction from his stock of mirrors and lenses. He would not have known that it took centuries to discover the wave nature of light (propagation, diffraction, interference, and coherence), its electromagnetic nature (polarization and propagation through anisotropic media), its quantum nature (photons and light–matter interaction), and to develop concepts such as thermal equilibrium, oscillation, modes, spectral analysis, and transient dynamics—all of which are necessary to truly understand, design, and use lasers.

4.2 The Laser: An Optical Oscillator

The laser is simply an oscillator of light, and the phenomenon of *oscillation* is one of its underlying foundational principles.

4.2.1 Oscillators

An oscillator is a device, system, or structure that produces oscillation at some frequency, with little or no excitation at that frequency. An example of a mechanical oscillator producing sound is the familiar tuning fork. When struck, it vibrates, or oscillates, creating sound at its characteristic (resonance) frequency (Fig. 4.1). The oscillation eventually decays since there is no energy source to sustain it. A musical instrument is made of mechanical structures (chords or pipes) that oscillate at distinct frequencies, which may be altered by changing their dimensions or shapes, as the instrument is played. Another example is the undesirable acousto-electrical oscillation often encountered when a microphone is connected to an audio amplifier feeding a close-by loudspeaker (Fig. 4.2). A small disturbance sensed by the microphone is amplified, and if the sound produced by the loudspeaker reaches the microphone, it gets amplified once more, and the cycle is repeated. Circulation through this feedback loop results in

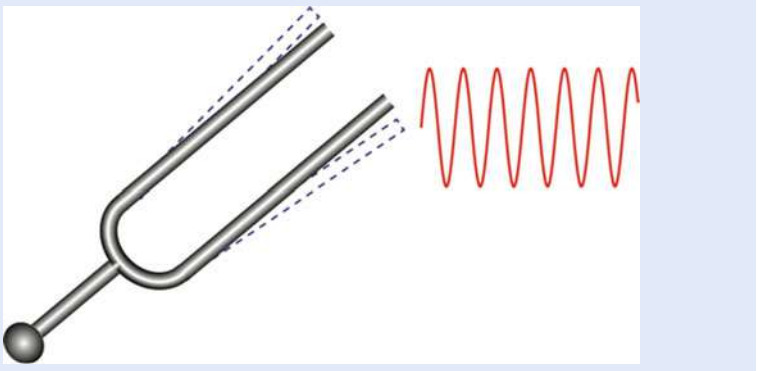


Fig. 4.1 When struck, a tuning fork vibrates at its characteristic resonance frequency

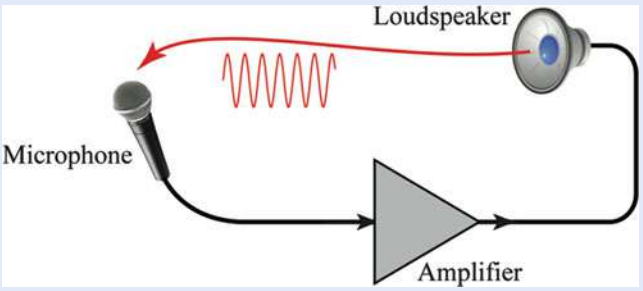


Fig. 4.2 Undesirable oscillation is created when the loudspeaker sound reaches the microphone

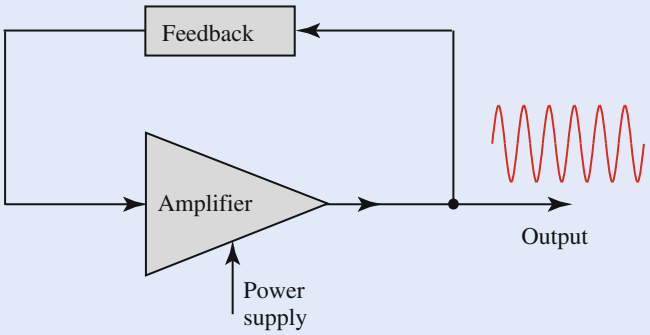


Fig. 4.3 An electronic oscillator using an amplifier and a feedback loop

oscillation—the familiar whistle sound. The frequency of oscillation is characteristic of the overall system.

The oscillator is a basic building block of virtually all electronic systems, analog or digital. An electronic oscillator comprises a resonant electronic amplifier along with a feedback system that directs the output of the amplifier back to its input, in the form of a feedback loop, as illustrated in Fig. 4.3. It is essential that the feedback be positive, i.e., the feedback signal re-entering the amplifier must be in phase with the original signal. Oscillation is initiated by noise, which contains a broad spectrum of all frequencies. The resonant amplifier amplifies a selected frequency component, and the feedback circuit brings the output back to the input for further amplification. For example, with gain of 2, a tiny input traveling 20 times through the feedback loop is amplified by a factor of $2^{20} \approx 10^6$. Of course, the output of the device cannot grow without bound, since the amplifier eventually saturates, i.e., its gain is reduced when its input becomes too large. As illustrated in Fig. 4.4, when the reduced gain eventually becomes equal to the loss encountered in the loop, growth of the circulating signal ceases, and the oscillator output stabilizes. The ultimate result is the generation of energy at a specific resonance frequency, with little initial excitation at that frequency. Energy is of course provided by the amplifier power supply, e.g., a battery. Resonant amplification is achieved by means of a resonant element, usually in the form of a capacitor connected to an inductor in the domain of electronics, whose values determine the resonance frequency.

High-frequency electronic oscillators were developed as electronics technology advanced. Electronics emerged in the first half of the twentieth century and has advanced steadily with efforts to achieve miniaturization and greater switching

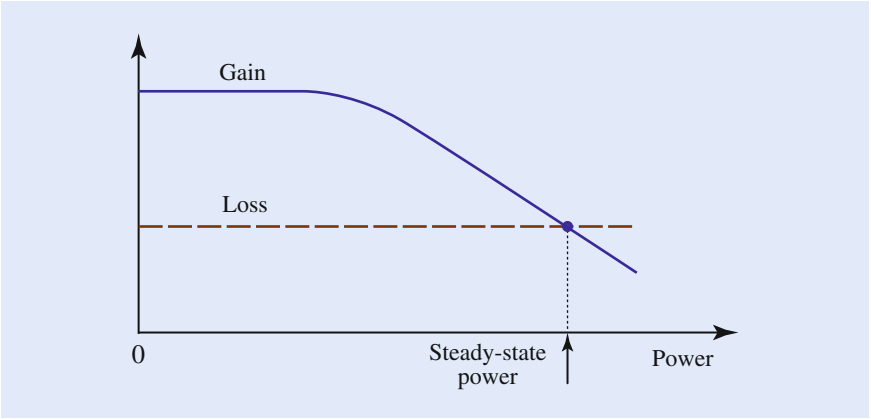


Fig. 4.4 Buildup of oscillation. As the power increases, the amplifier gain is reduced; when it equals the loss in the feedback loop, a steady-state power is reached

speeds—and this continues to this day. The quest to build electronic oscillators that operate at higher and higher frequencies was fueled by the desire to make use of wider and wider bands of the electromagnetic spectrum. The earliest electronic oscillators operated at audio frequencies (AF) and radio frequencies (RF), in the kHz and MHz ranges. They used transistor-based amplifiers and inductor–capacitor resonant circuits. Early microwave (MW) oscillators in the GHz range employed special vacuum-tube amplifiers, such as magnetrons and klystrons, which were based on ingenious mechanisms for forging interactions of the microwave field with electron beams, along with microwave cavity resonators that provided the prerequisite feedback. These oscillators were used for decades following their introduction in the 1940s, but were ultimately replaced by solid-state devices such as Gunn and IMPATT diodes.

The development of AF, RF, and MW oscillators was motivated by the needs of the electronics and telecommunications industries, and the development of oscillators at frequencies in the GHz range was fueled by the need for radar systems during the Second World War. This evolutionary process was bound to lead naturally to the development of oscillators in the THz frequency range and beyond, including optical frequencies (the frequency of visible light lies in the 400–770 THz range, and ultraviolet light and X-rays have even greater frequencies), but this natural evolution took several decades. This may be because the need for a new kind of light source was not pressing since many conventional light sources already existed. Gas-discharge lamps emitting over a narrow band of frequencies, filtered both temporally and spatially, met the needs of most scientific applications envisioned at the time.

Perhaps it was the development of the maser, the microwave predecessor of the laser, that stimulated the extrapolation to optical frequencies; both the laser and the maser are based on the same amplification principle: *stimulated emission*. In fact, after the laser was invented, it was known as an *optical maser* in the early technical literature. The term "maser" is an acronym for Microwave Amplification by Stimulated Emission of Radiation.

4.2.2 The Optical Oscillator

As mentioned earlier, two basic mechanisms are necessary to build an oscillator: resonant amplification and positive feedback. The frequency of oscillation is dictated not only by the resonant amplifier, but also by the feedback system,

4

because of the requirement that the phase of the feedback signal arrives in phase with the initial input. An energy source is of course necessary to support the amplification process. In essence, the device converts the power supply energy to energy at the oscillation frequency. How might these basic mechanisms be implemented in the optical regime in order to build an optical oscillator?

Before we go further, it is important to note that optical oscillators exist naturally and need not be invented! Every single excited atom emitting light is itself an optical oscillator. When the atom undergoes a transition between two energy levels with energy difference ΔE it produces a photon at frequency $\nu = \Delta E/h$, where h is Planck's constant. The transition may be regarded as an oscillator with resonance frequency ν. The problem is that these identical atoms emit photons independently, in random directions and with no common phase, although they do have the same resonance frequency. The acoustic analog of such an incoherent source is a large set of independently struck tuning forks, or a large orchestra playing without a conductor. The outcome, which is a superposition of light emitted by say 10^{23} independent atoms, is an optical field that oscillates randomly in both time and space. This is light that lacks temporal and spatial coherence, i.e., it is basically optical noise. What is needed instead is a *single* optical oscillator producing a coherent optical field, rather than an extensive set of independent oscillators, each with miniscule power. The laser does just that!

The laser is an optical oscillator that employs a *coherent* optical amplifier—a medium, which when illuminated by incoming light, produces more light in the same direction and with the same properties as the incoming light. Gordon Gould, one of the early pioneers of the laser, is credited with calling this mechanism Light Amplification by Stimulated Emission of Radiation, thereby introducing the acronym LASER. Since the laser is an oscillator, rather than an amplifier, a more appropriate name would perhaps be Light Oscillation by Stimulated Emission of Radiation, but the associated acronym would have been imprudent.

To construct a system that acts as a single oscillator, i.e., a coherent light source, feedback is necessary. With adequate feedback, the stimulated emission mechanism synchronizes the individual independent atomic oscillators to act collectively as a single optical oscillator. Feedback can be readily implemented by means of an arrangement of mirrors or reflective surfaces that form an optical resonator within which light is repeatedly passed through the amplifying (or gain) medium. The simplest optical resonator takes the form of two parallel planar or spherical mirrors between which the light circulates back and forth, as illustrated in Fig. 4.5. This structure, known in the optics community as a Fabry–Pérot interferometer, was not readily recognized as a resonator until the laser was invented. Amplification occurs at each passage since the gain medium amplifies in both directions. Useful light is extracted by making one of the mirrors partially transmitting.

Every oscillator has a built-in mechanism for stabilization that governs the steady output power (see Fig. 4.4). In lasers, if the pump power is sufficiently high such that the gain in the medium is greater than the resonator loss, then

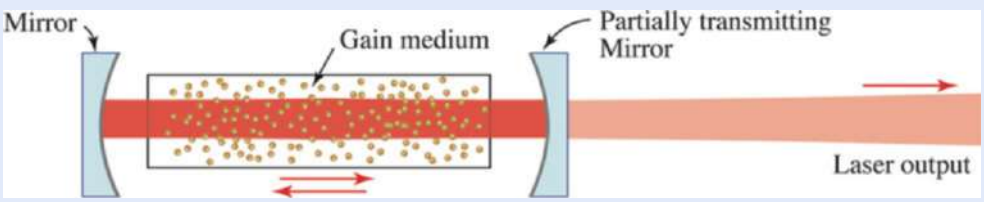


Fig. 4.5 The laser is an optical oscillator that makes use of an amplifying (gain) medium placed inside a resonator

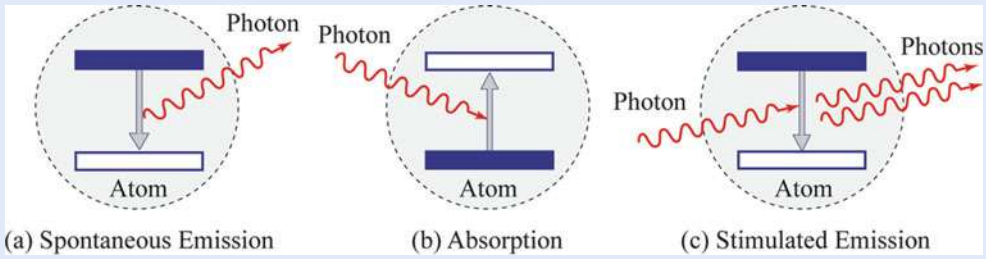


Fig. 4.6 Interaction of light with matter by transitions between two energy levels. (**a**) Spontaneous emission of a photon by an excited atom. (**b**) Absorption of a photon by an unexcited atom. (**c**) A photon stimulates an excited atom to emit a second photon

lasing is initiated and the optical power increases exponentially. However, since atoms undergoing stimulated emission become de-excited in the process, the population inversion is thereby reduced so that further growth of the optical power is suppressed. When the decreased gain eventually equals the loss, the system reaches equilibrium and a steady laser power is delivered.

4.2.3 Optical Amplification by Stimulated Emission

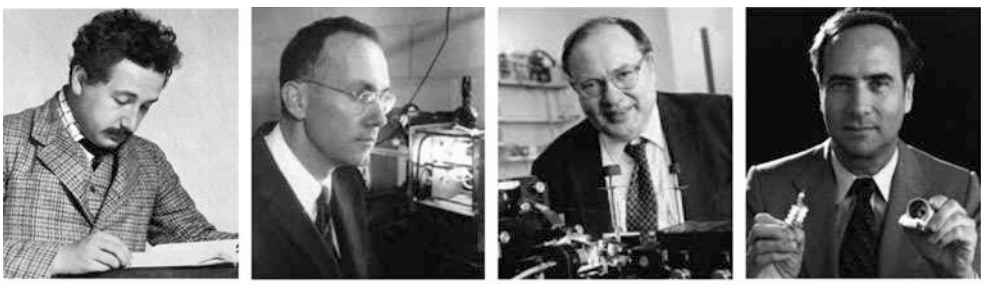
There are three processes inherent in the interaction between photons and atoms (see Fig. 4.6). The first is *spontaneous emission*, which does not generate coherent light. The second is *absorption*, which causes the medium to become an attenuator rather than an amplifier. When unexcited atoms absorb light, they are excited to the higher energy level and subsequently decay back spontaneously, either by emitting light or via some other non-radiative means. *Stimulated emission* is the third component of the interaction. The relation among these processes was set forth by Albert Einstein in 1917 when he revisited Max Planck's law of radiation (the spectral distribution of blackbody radiation in thermal equilibrium).

Einstein showed that while an atom in an unexcited state (lower energy level) might absorb a photon at a rate proportional to the incoming photon flux density, an atom in an excited state (higher energy level) is just as likely to be stimulated by the incoming photons and to emit a photon, so that the flux of photons increases. An important recognition was that the photon emitted via stimulated emission is in the same direction and has the same properties as the stimulating photon. These three mechanisms—absorption, and stimulated and spontaneous emission—govern the interaction of photons with atoms and underlie the law of radiation in thermal equilibrium as well as the generation of laser light.

For a medium with an excess of unexcited atoms, absorption exceeds stimulated emission and the medium provides net attenuation. This is the situation when the medium is in thermal equilibrium. If this equilibrium is somehow disturbed to create an excess of excited atoms, stimulated emission dominates absorption and the medium provides net gain. The medium then serves as a coherent optical amplifier. This non-equilibrium state, called *population inversion*, may be achieved via a process known as *pumping*. In fact, the pumping process supplies the medium with the energy needed to realize the desired gain. Since transitions occur only when the optical frequency matches the resonance frequency of the atomic transition (i.e., the energy of the photon $h\nu$ matches that of the atomic transition energy ΔE), the amplifier is a resonant amplifier. It provides gain only within a narrow spectral band dictated by the atomic transition.

Following Einstein's conception in 1917, the fact that stimulated emission could provide optical gain was confirmed in 1928 by Rudolf W. Ladenburg.

Its use for amplification (also called *negative absorption*) was predicted by Valentin A. Fabrikant in 1939. Optical pumping, i.e., achieving population inversion by the use of another light source, was proposed in 1950 by Alfred Kastler (Nobel Prize for Physics in 1966) as a mechanism for introducing gain. Despite these early discoveries of the basic ingredients of this coherent optical amplifier, it was not until 1960 that a laser was constructed and successfully operated by Maiman. The first maser had been built in 1953 by Charles H. Townes, James P. Gordon, and Herbert J. Zeiger. Townes, Nikolai G. Basov, and Aleksandr M. Prokhorov were awarded the 1964 Nobel Prize for Physics for theoretical work leading to the maser. Masers are now used as low-noise microwave amplifiers for applications such as radio telescopes. In 1958 Arthur L. Schawlow (Nobel Prize for Physics in 1981), together with Townes, suggested a method for extending the stimulated-emission principle of the maser to the optical region of the spectrum.



4.2.4 Laser Materials and Pumping Methods

An enormous variety of materials are used as gain media in lasers and laser amplifiers; these include solids, gases, liquids, and plasmas. The wavelengths of these devices span extended bands of the electromagnetic spectrum, all the way from the microwave to the X-ray region. They make use of a wide variety of resonator configurations and pumping schemes (Fig. 4.7). Pumping may be implemented optically by means of light at a wavelength other than the resonance wavelength, e.g., by a flash lamp or another laser; or it may be implemented by use of an electric current, as is the case in semiconductor laser diodes.

4.3 Optical Resonators and Their Modes

Optical feedback is commonly provided by use of a resonator, which serves as a "container" within which the generated laser light circulates and is built up, and can be stored. Optical resonators take a variety of configurations, as illustrated by the examples in Fig. 4.7. The most common configuration is two planar or spherical mirrors. Ring lasers use an arrangement of mirrors in a ring configuration or use a closed-loop optical fiber or integrated-optic waveguide. Dielectric waveguides with two cleaved end surfaces are used in semiconductor lasers. In microdisks and microspheres, light circulates inside the rim of the material by total internal reflection at near-grazing incidence, in what are known as whispering-gallery modes. Periodic dielectric structures such as distributed Bragg reflectors

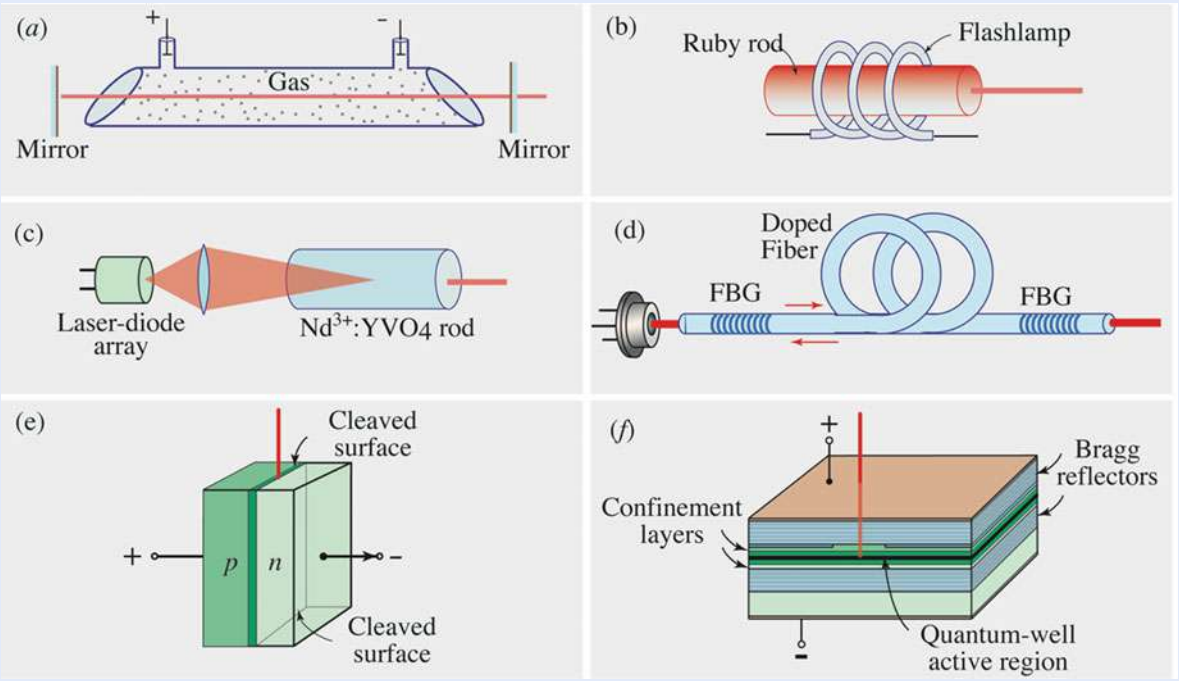


Fig. 4.7 Examples of lasers. (**a**) Gas laser pumped by direct current (DC). (**b**) Solid-state laser, e.g., ruby, optically pumped with a flashlamp. (**c**) Nd:YVO4 solid-state laser optically pumped by a laser diode array. (**d**) Fiber laser (e.g., erbium-doped silica fiber) with fiber Bragg grating (FBG) reflectors, pumped with a laser diode. (**e**) Laser diode (forward-biased *p–n* junction) with cleaved surfaces acting as mirrors, pumped by electric-current injection. (**f**) Quantum-well semiconductor laser, pumped electrically. Charge carriers are restricted to the active region by the confinement layers and Bragg reflectors serve as mirrors

(DBRs) are used as mirrors for trapping the light inside a fiber or in small structures, as in the micropillar resonator. Light can also be trapped in defects within dielectric photonic-bandgap structures, forming photonic-crystal resonators. In microresonators, which are used in microlasers, the size of the resonator can be of the same order of magnitude as its resonance wavelength. Nanoresonators, which can be far smaller than the resonance wavelength have come to the fore in recent years.

The optical resonators are characterized by their quality factor Q, which is the ratio of the resonance frequency to the line width. A low-loss resonator has a large Q, corresponding to sharp resonance and long storage time (in units of optical period). Resonators may have Q as high as 10^8, corresponding to a narrow spectral width of 3 MHz and a storage time of approximately 0.3 μs, for resonance at a wavelength of 1 μm.

4.3.1 Modes

Another foundational principle underlying the laser is that of *modes*. A resonator supports light in specific spatial and longitudinal modes. Modes are fields that self-reproduce as they circulate through the resonator. Spatial modes are spatial distributions that maintain their shape after one round trip. Longitudinal (or spectral) modes are fields with frequencies for which the circulating light arrives in the same phase (or shifted by multiples of 2π) after one round trip.

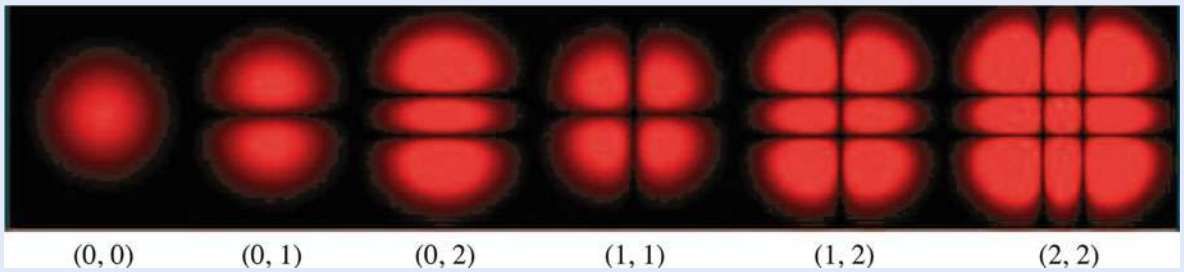


Fig. 4.8 Spatial distributions of the Hermite–Gauss modes of spherical-mirror resonators. The Gaussian mode, which is the lowest-order mode (0, 0), is the most confined around the resonator axis

As mentioned earlier, this ensures positive feedback, which is a necessary condition for oscillation. Each spatial mode may support multiple longitudinal modes.

Laser oscillations occur in those spatial and longitudinal modes for which the round-trip gain is greater than the loss. Since the gain is available within the spectral range defined by the atoms, only those modes whose frequencies lie in this range may oscillate. Moreover, since the spatial modes have different spatial profiles, and therefore undergo different losses, only a finite number of spatial modes oscillate, with the more confined modes favored. And each of these spatial or spectral modes has two polarization degrees of freedom, constituting polarization modes (horizontal/vertical linear polarization or right/left circular polarization).

Ideally, the laser is designed to operate in a single spatial mode with a single longitudinal mode. Such a single-frequency laser is a single optical oscillator with the highest spatial and temporal coherence. However, lasers are also often designed for operation in a single spatial mode with many longitudinal modes. Since these modes oscillate independently, their sum undergoes interference (beating) resulting in amplitude variation and a broader spectrum with reduced temporal coherence. However, such variations are usually on a short time scale, e.g., nanoseconds, so that the average power remains steady when averaged over a longer time, which serves well for certain applications. Operation in multiple spatial modes, each with its own longitudinal modes, can be useful for applications requiring greater power, although such multimode lasers exhibit reduced temporal and spatial coherence.

The spatial modes of the spherical-mirror resonator are the Hermite–Gauss modes illustrated in Fig. 4.8. The widths of these modes and their divergence angles are determined by the curvatures of the mirrors and their separation.

4.3.2 The Gaussian Beam

The Hermite–Gauss mode with the smallest width is the Gaussian mode. This is responsible for generating the Gaussian beam (Fig. 4.9), which has the smallest angle of divergence for a given width. Its diffraction-limited divergence angle is inversely proportional to the beam radius at its waist W_o, namely $\theta_o = \lambda/\pi W_o$. For example, for $\lambda = 1$ μm and $W_o = 1$ cm, $\theta_o = 3.18 \times 10^{-5}$ radians. At a distance $d = 3.8 \times 10^8$ m (the distance to the moon), this corresponds to a spot diameter $2d\theta_o \approx 24$ km. If the beam radius were increased from 1 cm to 1 m, the spot size at the moon would be only 240 m. The Gaussian laser beam is highly collimated near its waist so that it may be regarded as a planar wave and used for applications requiring plane waves.

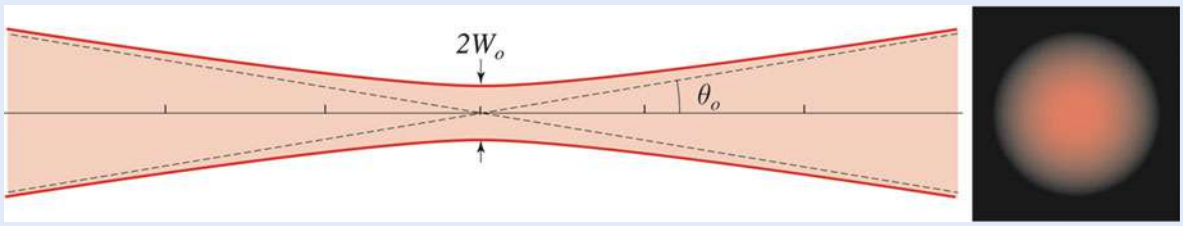


Fig. 4.9 The Gaussian beam

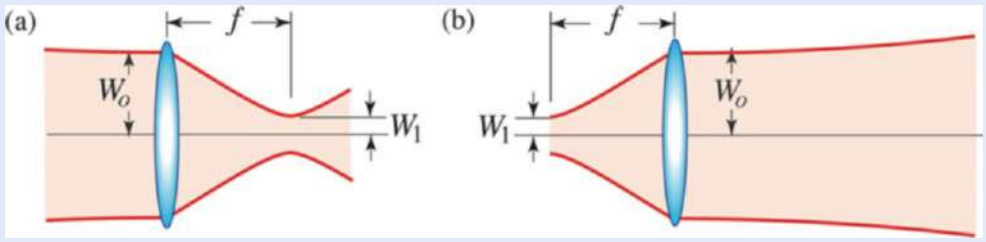


Fig. 4.10 Manipulation of a Gaussian beam by use of a lens of focal length *f*. (**a**) Reduction of the beam waist (focusing). (**b**) Reduction of the divergence angle (collimation)

The angle of divergence, and the associated waist, of the Gaussian beam may be manipulated by the use of lenses. For example, a beam with large waist W_o (and small divergence angle) may be used to generate a beam of smaller waist (and large divergence angle) W_1 by use of a lens of short focal length *f*. This is important for applications in lithography and laser scanning microscopy or for industrial applications such as cutting and welding. Conversely, a beam of small width and large divergence angle, such as that generated by a laser diode, may be converted into a beam with small divergence angle by use of a lens, as shown in Fig. 4.10. This is used in laser pointers.

4.4 Coherence of Laser Light

The most unique property of the laser is its temporal and spatial coherence. Laser light has a long coherence time (narrow spectrum) and a large coherence area. Conventional light sources have short coherence times (broad spectra) and small coherence area; they are incoherent.

The *coherence time* τ_c is the time duration over which the wave maintains its phase. This quantity is inversely proportional to the spectral width of the light. Laser light has a very narrow spectral distribution, ideally single frequency (a single wavelength), i.e., it is monochromatic or single color. Temporal coherence is also called longitudinal coherence. The *coherence length* $\ell_c = c\tau_c$, where *c* the speed of light. Long coherence length signifies that the phase of the wave is correlated over a long distance along its direction of propagation. Temporal coherence allows the production of ultrashort pulses of light, as short as a femtosecond or even in the attosecond regime.

Spatial (or transverse) coherence describes the correlation of light fluctuations in the transverse plane. The coherence area is the area within which the wave is correlated. Because of its spatial coherence, laser light in the Gaussian spatial mode

undergoes minimal spread as it travels, so that the beam has the least divergence (diffraction-limited), i.e., is highly directional, and the smallest spot size at great distances. This feature is important for various applications including free-space communication. Also, the laser beam can be focused to a spot of minimal width, so that it provides maximal irradiance, an important feature for various applications such as lithography and other industrial applications.

It is essential to note that the coherence properties of light from a conventional incoherent source can be improved by means of spectral and spatial filters. A spectral filter with narrow spectral width can enhance the temporal coherence by filtering out frequencies outside a narrow spectral band. A spatial filter, which may be constructed by sending the light through a pinhole, enhances spatial coherence. Light originating from a point is radiated as a spherical wave, which may be converted into a planar wave with full spatial coherence. Such enhancement of the coherence was in fact practiced before the invention of the laser, when coherent light was generated and used to demonstrate wave properties of light such as interference and diffraction, and to form thin optical beams. The difficulty with such enhancements is that much of the power of the original light was lost by the filtering process. An abiding advantage of laser light is that it combines excellent coherence properties with high power. This explains al-Hasan's frustration in the fictitious story about his attempt to convert light from a lantern into a beam resembling that from the laser. His effort to create order out of chaos ended up with discarding most of the light itself.

Can coherent laser light be distinguished from light generated by a thermal source that is filtered temporally and spatially to have the same coherence time and coherence area as the laser light? The answer is yes! Measurements of other statistical properties of the light intensity and the photons do reveal a difference. One measure is the intensity correlation, introduced by Robert Hanbury Brown and Richard Q. Twiss, fueled major advances in classical and quantum coherence theory. Light from a thermal source exhibits intensity fluctuations whose variance is equal to the squared mean, so that $g^{(2)} = \langle I^2 \rangle / \langle I \rangle^2 = 2$, where I is the intensity, whereas for laser light this variance is zero, which leads to $g^{(2)} = 1$. Photon counts obey Bose–Einstein statistics for thermal light, and Poisson statistics for coherent light from a laser. These counting probability distributions are distinctly different. For a mean number of photons $\langle n \rangle$ detected in a fixed time interval, the count variance is $\langle n \rangle + \langle n \rangle^2$ for thermal light, while it is only $\langle n \rangle$ for coherent laser light.

4.5 Pulsed Lasers

With steady pumping exceeding the threshold required for lasing, the laser output power remains constant over time and the laser is said to be *continuous wave* (CW). Pulsed lasers produce optical power in the form of pulses of certain duration and repetition rate. The creation of short pulses is the temporal equivalent of *focusing* of power in space. Key features of the pulsed laser are the pulse energy, the peak power, and the average power. Higher peak powers may be obtained for shorter pulses with the same pulse energy. Certain applications require high peak power, while others call for high pulse energy. Pulsed operation at a low repetition rate provides adequate time between pulses for the pump to build up a population inversion, thereby allowing the generation of pulses with high energy for the same average power.

Laser pulse durations may be as short as femtoseconds and can be compressed to attoseconds. Pulse-repetition rates extend from hours to more than 10^{11} pulses per second, while peak powers can reach 10 MW. Some gain media are suitable

only for use in pulsed lasers since CW operation would require pumping at a steady power so high that it could be impractical or result in excessive heat.

There are two principal schemes for pulsed-laser operation. The first exploits the transient dynamics of the laser system and the energy storage capability of the resonator by on–off switching of the gain, the loss, or the fraction of light extracted from the resonator. These are, respectively, called *gain switching*, *Q-switching*, and *cavity dumping*. The energy is stored during the off-time and released during the on-time. In the second scheme, called *mode locking*, the set of independently oscillating longitudinal modes of the laser are locked together to produce a single periodically pulsed oscillator. Both of these schemes may be used to generate short laser pulses with peak powers far greater than the constant power deliverable by CW lasers.

Gain switching is based on pulsing the pumping source. This is feasible if the pulsing time scale is much slower than time scales governing the lasing process. Examples of gain switching include lasers using electronically charged flashlamps, and semiconductor laser diodes in which the electric current used for pumping is itself pulsed.

Q-switching is loss switching. During the off-time the resonator loss is increased (by spoiling the resonator quality factor Q) using a modulated absorber inside the resonator. Because the pump continues to deliver constant power at all times, energy is stored in the atoms in the form of an accumulated population difference. When the losses are reduced during the on-times, the large accumulated population difference is released, generating an intense short optical pulse.

Cavity dumping is based on storing light in the resonator during the off-times, and releasing it during the on-times. During the off-time, the pump is operated at a constant rate and the generated light is stored in the resonator, which is not allowed to transmit and has negligible losses. The light is subsequently released, or "dumped," as a useful pulse by suddenly removing one of the mirrors altogether (e.g., by rotating it out of alignment), increasing its transmittance to 100 %. As the accumulated light leaves the resonator, the sudden increase in the loss arrests the oscillation, and the process is repeated, resulting in strong pulses of laser light.

Mode-locking. Mode locking is the most important of the various techniques for generating ultrashort laser pulses, from tens of picosecond to less than 10 fs. This is attained by locking the phases of the longitudinal modes together. Since the frequencies of these modes are equally separated, they behave like the Fourier components of a periodic function, and therefore form a periodic pulse train with period equal to the round-trip time between the resonator mirrors (Fig. 4.11). The coupling of the modes is achieved by periodically modulating the losses inside the resonator. The pulse width is determined by the number and profile of the

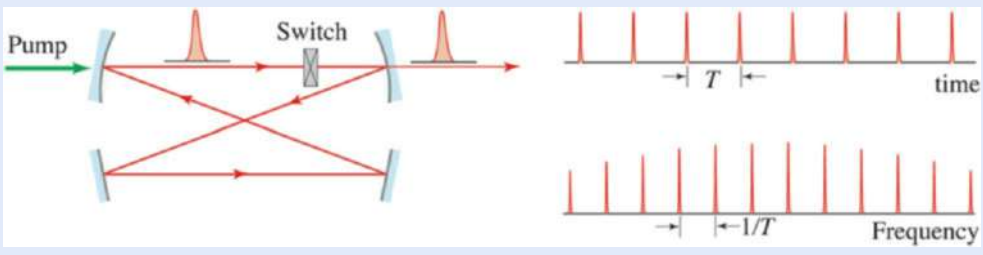


Fig. 4.11 The mode-locked laser. A pulse circulating inside a ring resonator periodically hits the exit mirror and is partially transmitted. The result is a pulse train with period equal to the round-trip time T. The spectrum of the emitted light, which is also periodic with period $1/T$, represents the now-locked longitudinal modes of the resonator. Locking is implemented by modulating the losses inside the resonator using a loss mechanism acting as a switch that lets a pulse out each time period T

spectral components (the modes). In accordance with Fourier theory, the wider the overall spectral width (i.e., the larger the number of modes), the shorter the pulse duration. For example, because of their wide spectral width, Ti:sapphire lasers generate pulses of only a few femtoseconds duration.

4.6 Conclusion

The history of the laser has been both an evolution and a revolution. As described earlier in this article, the development of the laser was an *evolutionary* process that drew together concepts formulated over a period of more than four decades (1917–1960). It also benefitted from the evolution of electromagnetic oscillators with increasing frequency, culminating in the maser. This evolution continues today, with new lasers of ever higher frequencies (wavelengths as short as tens of nanometers for X-ray lasers), pulses of durations shorter than 100 as, and greater optical powers: more than 100 kW for CW operation and peak powers in the GW regime (and as high as PW for laser fusion applications). Focused intensities can be as high as 10^{23} W/m^2.

There is no doubt, however, that within a few years of its invention the laser started a *revolution* in optics and optics-based technologies and applications, which created an abundance of opportunities in the early 1960s for major new scientific discoveries and a proliferation of novel technologies with far reaching applications.

Examples of the new sciences are: nonlinear optics, optoelectronics, laser spectroscopy, femtosecond physics and chemistry, attosecond atomic physics, and quantum optics. Laser applications have expanded to cover all aspects of modern technology:

Applications requiring high power and precision focusing include drilling, cutting, welding, ablation, material deposition, additive manufacturing, directed energy, and fusion.

Applications using the directional precision and focusing capability of the laser include optical disk drives, printers and scanners, barcode scanners, metrology and surveying, lithography, scanning microscopy, and adaptive optics imaging.

Applications utilizing high-speed modulation and switching include fiber-optic and free-space optical communications, optical cables, and interconnects.

Applications to medicine and health care include laser surgery, skin treatments, ophthalmic and cardiovascular diagnostics, laser vision correction (Lasik), and other diagnostic and therapeutic procedures.

The laser is truly a light fantastic!

Acknowledgments Discussions with M. C. Teich are acknowledged. Several of the figures used in this chapter were adapted from figures that appeared in *Fundamentals of Photonics* (2007) by B. E. A. Saleh and M. C. Teich.

Further Reading

1. Townes CH (1999) How the laser happened: adventures of a scientist. Oxford University Press, New York, paperback ed. 2002
2. Maiman T (2000) The laser odyssey. Laser, Blaine, WA
3. Kastler A (1985) Birth of the maser and laser. Nature 316:307–309
4. Bertolotti M (2004) The history of the laser. Taylor & Francis, London
5. Hecht J (2005) Beam: the race to make the laser. Oxford University Press, New York
6. Siegman AE (1986) Lasers. University Science Books, Mill Valley, CA
7. Silfvast WT (2008) Laser fundamentals, 2nd edn. Cambridge University Press, Cambridge
8. Saleh BEA, Teich MC (2007) Fundamentals of photonics, 2nd edn. Wiley, Hoboken, NJ
9. Svelto O (2010) Principles of lasers, 5th edn. Springer, New York
10. Milonni PW, Eberly JH (2010) Laser physics, 2nd edn. Wiley, New York

Solid-State Lighting Based on Light Emitting Diode Technology

Dandan Zhu and Colin J. Humphreys

D. Zhu • C.J. Humphreys (✉)
Department of Materials Science and Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge CB30FS, UK
e-mail: colin.humphreys@msm.cam.ac.uk

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_5

5

5.1 Historical Development of LEDs

More than 100 years ago in 1907, an Englishman named Henry Joseph Round discovered that inorganic materials could light up when an electric current flowed through. In the next decades, Russian physicist Oleg Lossew and French physicist Georges Destriau studied this phenomenon in great detail and the term 'electro-luminescence' was invented to describe this. In 1962, inorganic materials (GaAsP) emitting red light were first demonstrated by Holonyak and Bevacqua [1] at General Electric's Solid-State Device Research Laboratory in Syracuse, New York, although the light emitted was so weak that it could only be seen in a darkened room (by comparison, the efficacy of Thomas Edison's first incandescent light bulb was 10 times greater). Since then, the efficiency of GaP and GaAsP advanced significantly in the 1960s and 1970s. The AlInGaP system was developed later, in the 1980s, and is now the basis of most high-efficiency LEDs emitting in the red-to-yellow visible region. The development of the nitride material system (GaN, InN, AlN and their alloys) in the last two decades has enabled efficient light emission to expand into the blue and green spectral region, and most importantly, allowing the production of white light (blue is the high-energy end of the visible spectrum and therefore enables the production of white light using blue light plus phosphors). Blue LEDs were made possible by a series of key breakthroughs in materials science summarised in Table 5.1, which will be discussed in greater detail later. In particular, the first bright blue LED was announced at a press conference on November 12, 1993 by Nakamura [2]. The invention of efficient blue LEDs has enabled white light source for illumination. In 1997, white light was demonstrated for the first time by combining a blue gallium nitride (GaN) LED with a yellow-emitting phosphor [3]. Such LEDs are called 'white LEDs'.

Nowadays, solid-state lighting based on LEDs is already commercialised and widely used, for example, as traffic signals, large outdoor displays, interior and exterior lighting in aircraft, cars and buses, as bulbs in flash lights and as backlighting for cell phones and liquid-crystal displays. With the continuous improvement in performance and cost reduction in the last decades, solid-state

Table 5.1 A summary of the key steps in GaN-based LED development history

1938	Juza and Hahn [84]	The earliest polycrystalline GaN powder was synthesised by reacting ammonia with liquid Ga metal
1969	Maruska and Tietjen [92]	First single crystal GaN film was grown by chemical vapour deposition directly on a sapphire substrate
1972	Pankove et al. [102]	First blue GaN metal-insulator-semicondutor LED was reported
1986	Amano et al. [79]	Crack-fee GaN films with good surface morphology and crystallinity were achieved by growing a thin AlN buffer deposited on sapphire at low temperature before GaN growth
1989	Amano et al. [43]	Amano, Akasaki and co-workers demonstrated that a low-energy electron beam irradiation treatment in a scanning electron microscope could cause a previously highly resistive Mg-doped GaN layer to show distinct *p*-type conductivity, enabling the first GaN *p*–*n* junction LED
1991	Nakamura et al. [38, 94]	Nakamura and co-workers showed that a ~20 nm thick GaN buffer layer deposited at low temperature (~500 °C) before the main GaN growth at ~1000 °C could also be used to grow smooth films on sapphire, including *p*-type material with good electrical properties
1992	Nakamura et al. [42]	Thermal activation of Mg-doped GaN to achieve *p*-type conductivity
1993	Nakamura et al. [97]	Blue and violet emitting double-heterostructure (DH) LEDs were successfully fabricated
1993	Nakamura et al. [2]	Nakamura announced the first bright blue LED at a press conference on November 12, 1993
1995	Nakamura et al. [95]	InGaN quantum well LEDs were fabricated
1997	Nakamura et al. [3]	White light was demonstrated for the first time by combining a blue gallium nitride (GaN) LED with a yellow-emitting phosphor

lighting has emerged to be a realistic replacement of incandescent and fluorescent lamps for our homes and offices.

Compared with any other existing lighting technology, solid-state lighting possesses two highly desirable features: (1) it is highly energy efficient with tremendous potential for energy saving and reduction in carbon emissions; (2) it is an extremely versatile light source with many controllable properties including the emission spectrum, direction, colour temperature, modulation and polarisation. The beneficial impact of LEDs on the economy, environment and our quality of life is so evident and well recognised that the 2014 Nobel Prize in Physics was awarded to the inventors of efficient blue LEDs: Isamu Akasaki, Hiroshi Amano and Shuji Nakamura.

5.2 The Importance of Nitride Materials

The main compound semiconductor materials used in LEDs and their bandgap energies are summarised in Fig. 5.1. For most optoelectronic devices such as light emitting diodes (LEDs), laser diodes, and photodetectors, a direct bandgap is essential for efficient device operation. This is because the optical emission processes in a semiconductor with an indirect bandgap require phonons for momentum conservation. The involvement of the phonon makes this radiative process much less likely to occur in a given timespan, which allows non-radiative processes to effectively compete, generating heat rather than light. Therefore semiconductors with an indirect bandgap are not suitable for efficient LEDs.

Conventional cubic III–V compound semiconductors, such as the arsenides and phosphides, show a direct-to-indirect bandgap transition towards higher energies. Therefore high-efficiency devices can be achieved in the infrared and red-to-yellow visible spectral regions, but the efficiency decreases drastically for

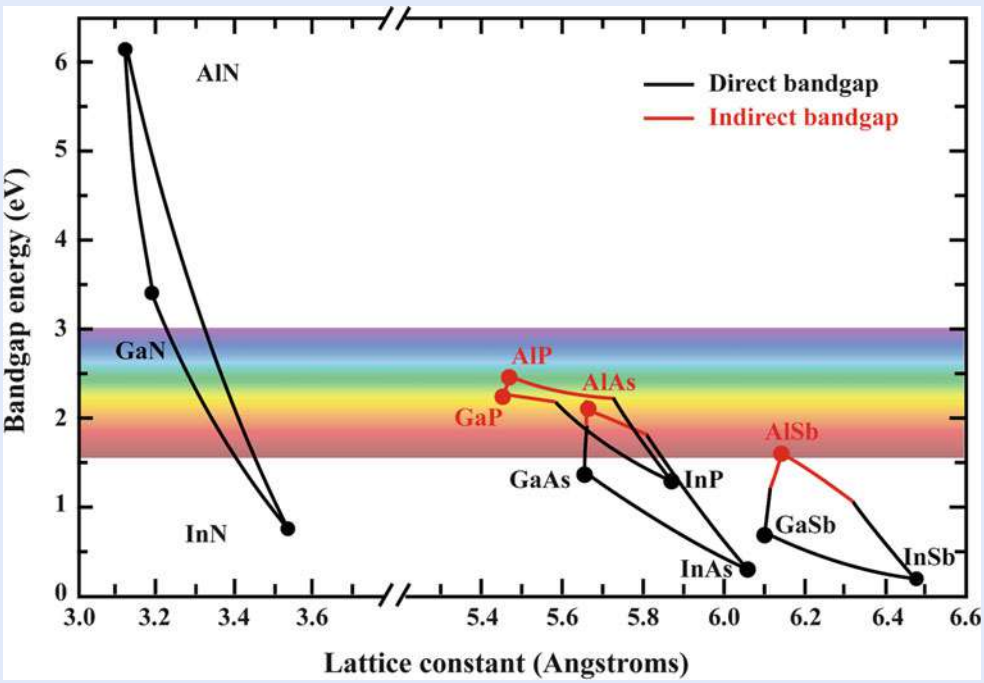


Fig. 5.1 Bandgap energies at 300 K of III–V compound semiconductors, plotted from data given in Vurgaftman et al. [4] and Vurgaftman and Meyer [5]. For the nitrides, the hexagonal *a* lattice constant has been used. The energy range corresponding to the visible spectrum is also indicated

conventional III–V semiconductors as the bandgap becomes indirect. In contrast, the nitrides have the hexagonal wurtzite structure, and the bandgap remains direct across the entire composition range from AlN to InN, with the bandgap energy covering a wide range from the deep ultraviolet to the infrared region of the electromagnetic spectrum. This makes the group-III nitrides system (consisting of GaN and its alloys with Al and In) particularly suitable for LEDs.

The blue/green and near-UV spectral regions can be accessed using the InGaN alloy, and today, the main application of the nitrides is in blue, green and white emitting LEDs, as well as violet laser diodes used for high-density optical storage in Blu-ray DVDs [6]. Since the InGaN bandgap energy spans the visible spectrum, extending into the infrared to ~0.7 eV for InN, this alloy covers almost the entire solar spectrum, and is thus a potential system for high-efficiency multi-junction solar cells [7].

The wide bandgap of the AlGaN alloy system will enable the fabrication of UV emitters and photodetectors. Possible applications of UV optoelectronics include water purification, pollution monitoring, UV astronomy, chemical/biological reagent detection and flame detection [8, 103].

AlGaN/GaN heterostructures are also suitable for electronic devices such as high electron mobility transistors (HEMTs), which have applications in microwave and radio frequency power amplifiers used for communications technology [9]. Such a wide bandgap materials system also allows device operation at higher voltages and temperatures compared to conventional Si, GaAs or InP-based electronics [10].

Although this chapter will be mainly focused on nitride-based LEDs for lighting applications, it is worth bearing in mind the great potential of nitride materials in other exciting applications mentioned above. And because of their unique materials properties and wide range of applications, group-III nitrides are widely considered to be the most important semiconductor materials since Si.

5.3 LED Basics

The simplest LED structure is a *p*–*n* junction, consisting of a layer of *p*-type doped semiconductor material connected to an *n*-type doped layer to form a diode with a thin active region at the junction. The principle for light emission in a *p*–*n* junction is illustrated in Fig. 5.2. The *n*-type region is rich in negatively charged electrons, while the *p*-type region is rich in positively charged holes. When a voltage is applied to the junction (called forward bias), the electrons are injected from the *n*-type region and holes injected from the *p*-type region across the junction. When the electrons and holes subsequently meet and recombine radiatively, the energy released is given out as light with an emission wavelength close to the bandgap of the material incorporated in the active region around the junction. For high efficiency, a heterojunction (consisting of two semiconductor materials with different bandgap) is usually preferred to a homojunction (consisting of a single semiconductor material) due to better carrier confinement, as shown in Fig. 5.2c, i.e. the electrons and holes are spatially confined together in the active region with lower bandgap energy, which increase the chance of radiative recombination to produce light.

For most high-efficiency LEDs, quantum wells (QWs) are routinely used in the active region, which provide additional carrier confinement in one direction, improving the radiative efficiency, i.e. the internal quantum efficiency (IQE). Quantum wells consist of a very thin (few nm thick) layer of a lower bandgap material, such as InGaN, between higher bandgap barriers, such as GaN (see Fig. 5.3). The QW active region is sandwiched between two thicker layers of *n*-type doped and *p*-type doped GaN for electron and hole injection, respectively.

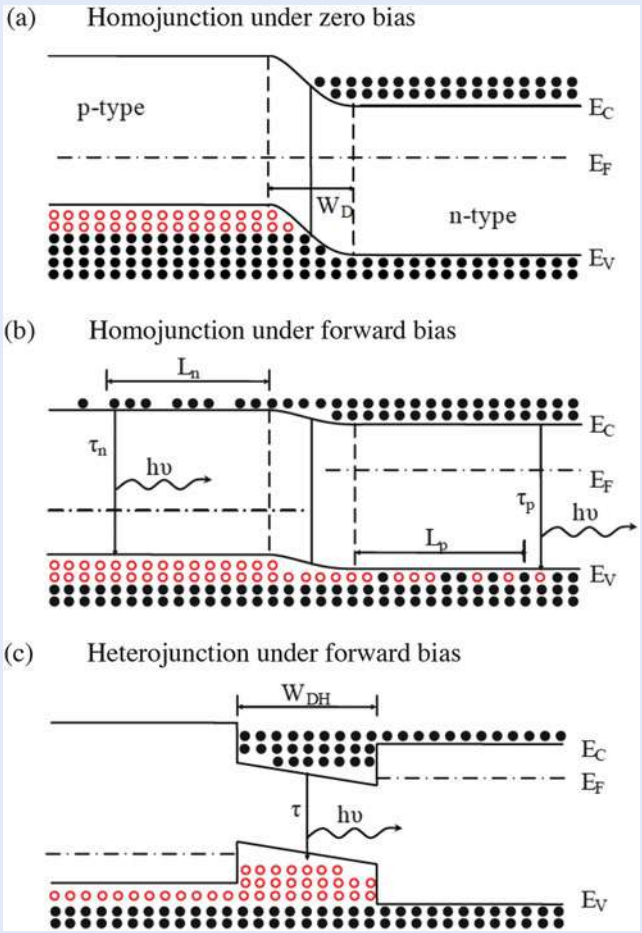


Fig. 5.2 A *p–n* homojunction under (**a**) zero and (**b**) forward bias. A *p–n* heterojunction under (**c**) forward bias. E_C, E_F and E_v are the conduction band, Fermi and valence band energy. *Filled circle* and *open circle* represent electrons and holes, respectively. In homojunctions, carriers diffuse, on average, over the diffusion lengths L_n and L_p before recombination. In heterojunctions, carriers are confined by the heterojunction barriers (after [11])

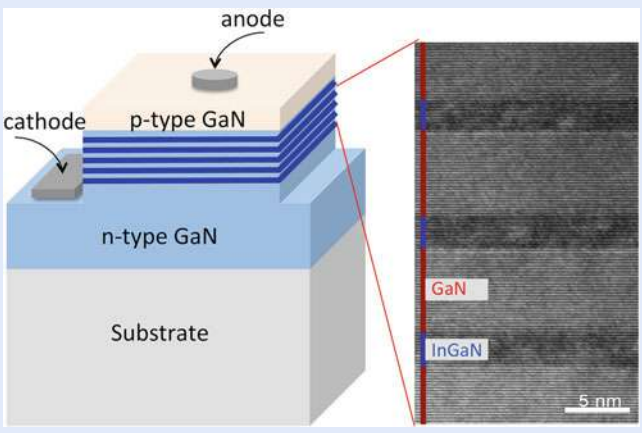


Fig. 5.3 A schematic InGaN/GaN quantum well LED structure together with a high-resolution transmission electron microscope lattice fringe image of three InGaN quantum wells separated by GaN barriers

The recombination of electron and holes across the InGaN quantum well region results in the emission of light of a single colour, such as green or blue. We can change this colour by varying the composition and/or changing the thickness of the InGaN quantum well.

5.4 Fabrication of an LED Luminaire

The LED structure described above is the essential source of light, but it often makes up only a tiny volume fraction of the final application, such as an LED light bulb or luminaire. Figure 5.4 illustrates the fabrication procedures involved in making an LED luminaire. The first step is the deposition of the nitride LED structure on a suitable substrate wafer such as sapphire, SiC, Si or GaN. This is performed by crystal growth usually via a process called metal organic vapour phase epitaxy (MOVPE) in a heated chamber or reactor. After deposition, these epiwafers will be processed into LED devices according to the LED chip design, which usually involves several steps including wafer bonding, *n* and *p*-type contact patterning, etching, metallisation and surface roughening. The processed LED devices are then separated via cleaving, sawing or laser cutting into individual dies. Depending on the target applications, these individual LED dies are mounted on an appropriate package in a form compatible with other electronic components such as drivers. For white LEDs, phosphors will also be incorporated into the package, together with blue-emitting LED dies in most cases. These packaged LED devices are then ready to be used as the light source in a luminaire.

From the fabrication procedure, we can see that there are many components contributing to the overall efficiency of a packaged LED device. These can be broken down into:

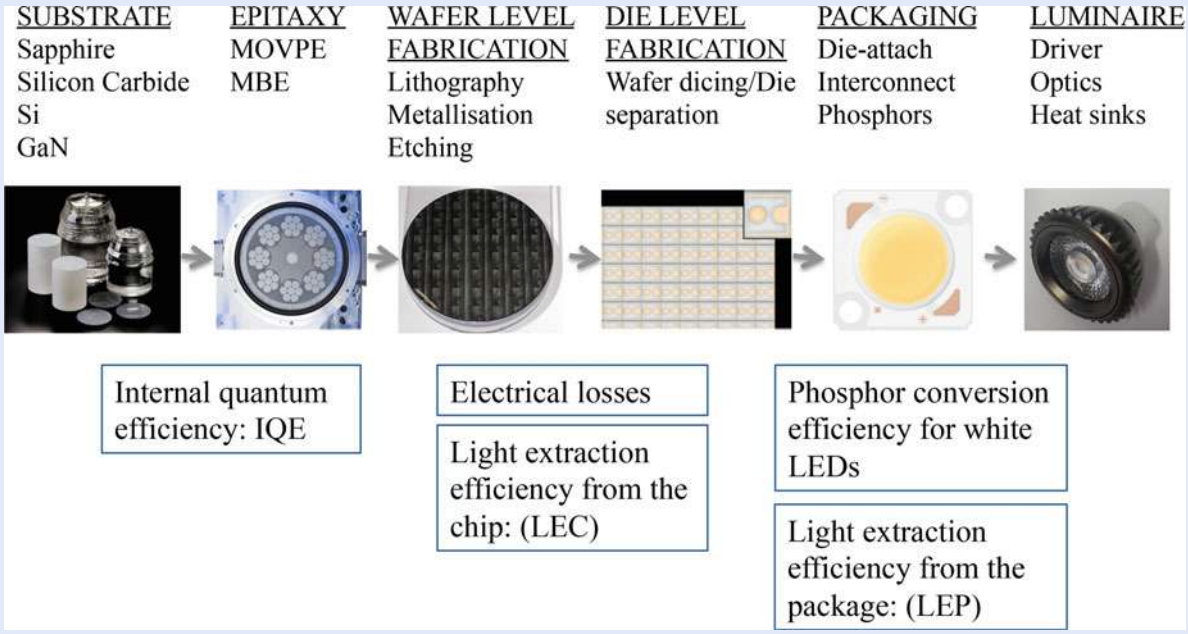


Fig. 5.4 Illustration of the fabrication procedures involved in making LED luminaries. The corresponding efficiency and losses involved in each procedure are also listed

1. Internal quantum efficiency (η_{IQE})
2. Light extraction efficiency from the chip (η_{LEC})
3. Electrical efficiency (η_{EE})
4. Phosphor conversion efficiency (η_{conv})
5. Light extraction efficiency from the package (η_{LEP})

The IQE is defined as the number of photons emitted from the active region divided by the number of electrons injected into the active region. The IQE is primarily determined by the LED structure design, such as the choice of material compositions, layer thicknesses, doping profile; and for a given structure, the material quality linked to the growth conditions used during the epitaxy procedure. The IQE is also a function of the current density through the LED. At high current density the IQE falls, a phenomenon known as 'efficiency droop'.

The light generated in the quantum well region needs to be extracted from the semiconductor material: most III–V semiconductors have high optical refractive indices (GaN: $n \sim 2.4$; InGaP: $n \sim 3.5$), and only a small portion of the light generated in the quantum well region can escape. This is because much of the light is trapped inside the LED by total internal reflection. Various advanced chip designs have been developed and used during the wafer and die level fabrication procedures to increase the possibility of light extraction from LED chips (LEC) and to minimise the electrical losses caused by the electrical contact and series resistances. Today, an LEC value >85 % is achieved for high performance commercial LED devices with a ThinGaN chip structure, as shown in Fig. 5.5b [12].

Furthermore LED dies need to be packaged before they can be incorporated with other electronic components in a real application. LED packaging is also critical to achieve high luminous efficiency, dissipate heat generated from the LED chip, improve reliability and lifetime and control the colour for specific requirements, as well as to protect the LED chips from damages due to electrostatic discharge, moisture, high temperature and chemical oxidation. A schematic structure of a high power LED package is shown in Fig. 5.5a, together with a picture of a commercial white LED package shown in Fig. 5.5c. The light extraction efficiency from a package (LEP) such as this is as high as 95 %. For white light generation, a yellow-emitting cerium-doped yttrium aluminium garnet (YAG) phosphor plate is added on top of the nGaN layer. To achieve a high phosphor conversion efficiency, the phosphor material is carefully chosen to match the LED emission for optimum excitation.

5.4.1 Efficiency and Efficacy

For a single colour LED such as blue, green and red LEDs, wall-plug efficiency is usually used as a measure of the overall efficiency. The wall-plug efficiency, measured by the light output power (measured in watts) divided by the electrical input (also in watts), is dimensionless and is usually expressed as a percentage. For white LEDs, a different term, efficacy, is usually used instead of efficiency. The unit of efficacy is lumens per watt (lm/W), corresponding to light power output (as perceived by the human eye and measured in lumens) relative to electrical power input (measured in watts). The terms efficiency and efficacy are both widely used in lighting, and care must be taken not to confuse them. The efficacy of a white light source will be explained in more detail later in this chapter. The term efficacy takes into account the sensitivity of the human eye to different colours: it is a maximum for green light at 555 nm.

It should also be noted that the efficiency or efficacy of a luminaire would be lower than the packaged LED devices due to additional losses caused by other

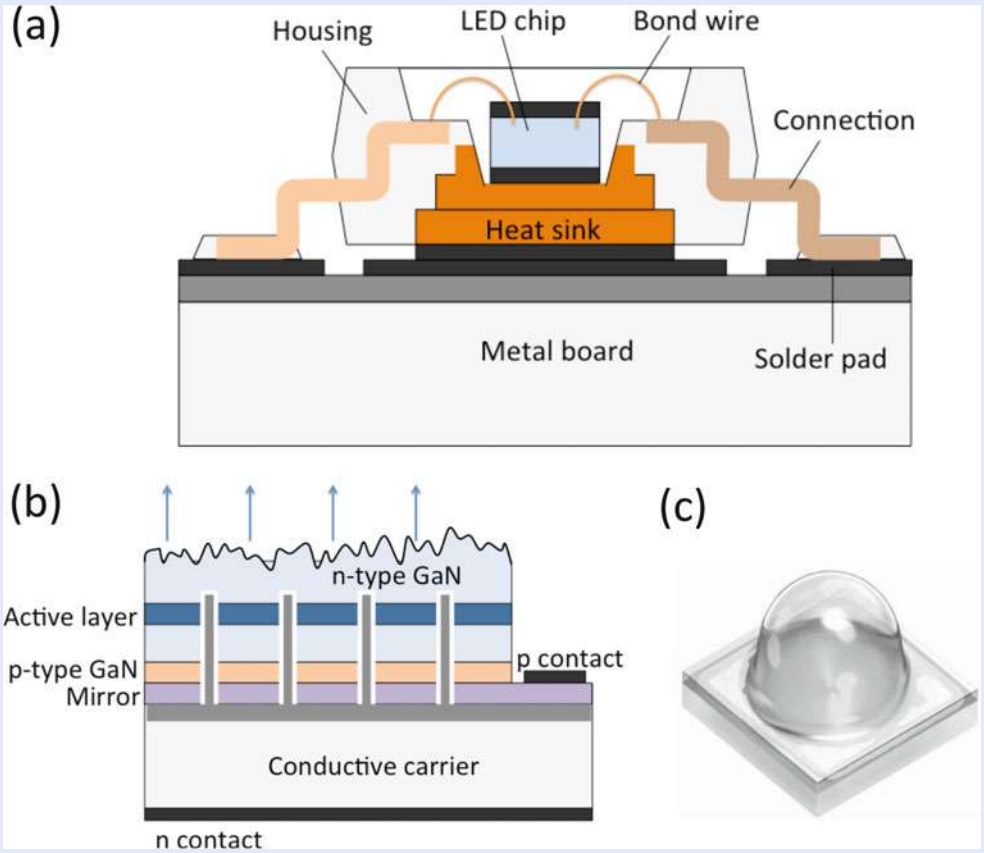


Fig. 5.5 (**a**) The schematic structure of a high-power LED package with good optical efficiency and thermal management, as required for high power LED chips. (**b**) Cross-section of a high power ThinGaN LED chip, illustrating the complex structure of state-of-the-art white LEDs for illumination. (**c**) A picture of a high power white LED package from Osram

components such as optics, heat sinks and electrical drivers. When discussing the efficiency of LED lighting, it is important to be clear about the form of the light source: whether it is a bare die, packaged LED device or luminaire.

The performance of LEDs has improved dramatically over the last decade with sustained improvements in the material quality, LED structure, chip design and packaging. Before moving to the discussions on LED performance and applications, it is worthwhile to first review the historical development of nitride LEDs, in particular the research challenges involved.

5.5 Research Challenges

The research in nitride materials and LED devices is a very broad and interdisciplinary field, spanning crystal growth, physics, materials science and characterisation, device processing, device physics, luminaire design and others. From a materials science point of view, nitride materials are highly defective compared with conventional semiconductor materials such as Si and GaAs, and the remarkable success of nitride-based LEDs is based on a series of wonderful achievements in science and engineering.

5.5.1 Crystal Growth

As with many other semiconductor materials, III-nitrides do not exist naturally, so the crystals need to be grown by some chemical reaction. The predominant growth method for the group-III nitrides is metalorganic vapour phase epitaxy (MOVPE, also called metalorganic chemical vapour deposition, MOCVD), both for research and mass-production of devices such as LEDs and lasers.

It should be noted that one key difference between the nitrides and the other III–V compound semiconductors mentioned earlier in this chapter is the lack of a suitable substrate for heteroepitaxial growth (namely, crystal growth on a different substrate material) of GaN. Bulk substrates of GaAs, GaP and InP can be used for epitaxy of most of the III–Vs and even II–VI compounds. Unfortunately, the nitrides have very high melting temperatures and dissociation pressures at melting, ~2800 K and ~40 kbar, respectively, for GaN, which means that bulk crystals cannot be grown from stoichiometric melts using the usual Czochralski or Bridgman methods [13,14]. Not only have bulk substrates of GaN been unavailable in a sufficient size and at reasonable cost, there is also no other suitable substrate material with a close lattice match to GaN. The properties of the GaN epitaxial layer such as crystal orientation, defect density, strain and surface morphology are to a large extent determined by the substrates used. Most commercial GaN-based LEDs are grown on sapphire or silicon carbide (SiC) substrates. Recently, the use of large area Si substrates has attracted great interest because high quality Si wafers are readily available in large diameters at low cost [106]. In addition, such wafers are compatible with existing sophisticated automated processing lines for 6 inch and larger wafers commonly used in the electronics industry.

Sapphire was the original substrate material, and remains the most commonly used to this day, but it has a lattice mismatch of 16 % with GaN. This is so large that attempts at direct epitaxial growth inevitably result in rough surface morphologies and a very high density of defects called dislocations that thread up through the growing layer: a typical density of such dislocations passing through the active InGaN quantum well region is five billion per square centimetre (5×10^9 cm^{-2}), as shown in Fig. 5.6.

The development of growth techniques for the reduction of the threading dislocation (TD) density in GaN on sapphire has resulted in considerable improvements. There are numerous methods in the literature, mostly related to the annealing of a low temperature nucleation layer [15], island formation and subsequent coalescence, as detailed in Figge et al. [16] and Kappers et al. [17,18]. An example of TD reduction using an SiN_x interlayer is shown in Fig. 5.7. The mechanism by which TD density can be reduced is as follows: the thin SiN_x interlayer constitutes a mask containing random holes through which small facetted GaN islands form on regrowth; aided by the inclined facets of the islands, the TDs bend laterally and react with other dislocations to annihilate and form half loops, hence halting their upward propagation, as illustrated in Fig. 5.7a. It was also found that the growth conditions of the GaN regrowth on top of the SiN_x interlayer have a pronounced effect on the degree of the TD reduction. By using a special 'slow' coalescence method, the TD density of the seed layer (5×10^9 cm^{-2}) was reduced to 5×10^8 cm^{-2} and successively deployed SiN_x interlayers reduce the TD density further to 1×10^8 cm^{-2}, as shown in Fig. 5.7b.

Dislocations are known to be non-radiative recombination centres [19] that should strongly quench light emission. Indeed, if the dislocation density in other semiconductors, for example, GaAs, exceeds around 1000 per square centimetre (10^3 cm^{-2}), the operation of light emitting devices is effectively killed. However, commercial InGaN blue and white LEDs show high performance despite the fact

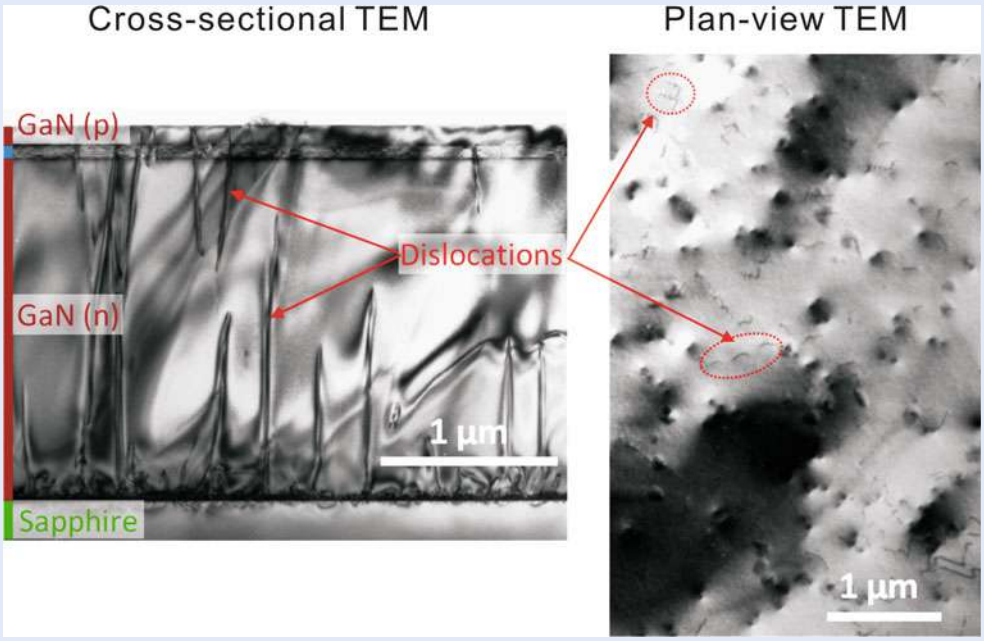


Fig. 5.6 Transmission electron microscopy (TEM) images showing the high density of threading dislocations resulting from the growth of GaN on sapphire substrate. The lattice mismatch between GaN and (0001) sapphire is 16 %, which gives rise to a dislocation density in the GaN of typically 5×10^9 cm^{-2}, unless dislocation reduction methods are used

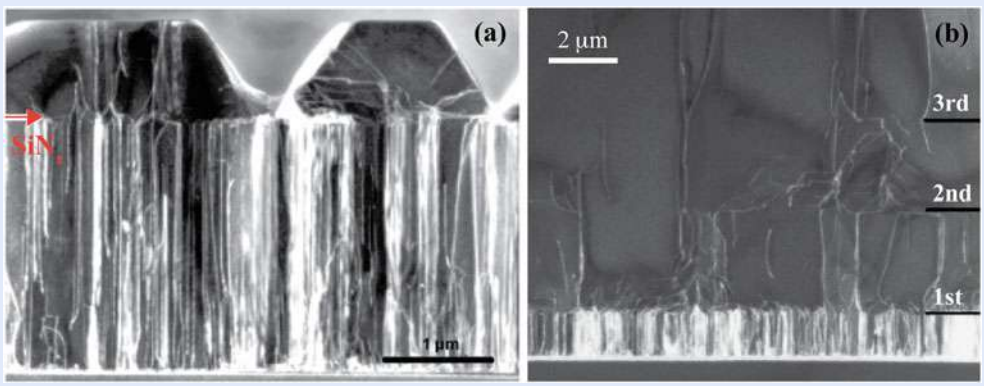


Fig. 5.7 (**a**) Cross-sectional TEM image of an SiN_x interlayer (*arrowed*) deposited on a GaN seed layer followed by the regrowth of GaN islands. Threading dislocations can be observed as bright lines in the image. (**b**) Weak beam dark field TEM image, $g = (11–20)$, showing the reduction of edge and mixed TDs with successive SiNx interlayers and a 'slow' coalescence of GaN between the layers

that the TD density of such devices is usually in the range of 10^8 cm^{-2}. The reason that InGaN LEDs are much more tolerant of TDs than other conventional III–V materials is probably due to carrier localisation effects [20–26]. The first contributing factor is the monolayer height interface steps on the InGaN quantum wells. Since the QWs are strained and because of the high piezoelectric effect in GaN, a monolayer interface step produces an additional carrier-confinement energy of about $2k_BT$ at room temperature, where k_B is the Boltzmann constant and T is the temperature. This is sufficient to localise the electrons. Recent three-dimensional atom-probe studies also confirmed that InGaN is a random alloy. Calculations show that random alloy fluctuations on a nanometer scale strongly

localise the holes at room temperature. Thus, the above two mechanisms can localise both the electrons and the holes, reducing diffusion to non-radiative defects like TDs. It is interesting to note that the electrons and holes are localised by different mechanisms in InGaN quantum wells.

Although high threading dislocation densities seem to be not very detrimental for InGaN LEDs, laser diodes and AlGaN-based UV-emitters do show a strong dependence of lifetime on dislocation density. Moreover, the growth conditions will also affect many microstructural properties of nitride materials as well as impurity levels and thus the final device properties. Therefore, the research in crystal growth remains highly relevant and important for high performance devices.

5.5.2 Internal Electric Field

The nitrides normally crystallise in the hexagonal wurtzite structure, which is non-centrosymmetric and has a unique or polar axis along a certain direction (the *c*-axis). Since the bonding is partially ionic due to the difference in electronegativity of the group III and V atoms, a spontaneous polarisation will exist in the crystal because of the lack of symmetry. In addition, most nitride devices involve the use of strained heterojunctions, such as InGaN/GaN. Because the in-plane lattice constant of InGaN is larger than for GaN, the InGaN layer will be under compressive strain perpendicular to the *c*-axis and under tensile strain along the *c*-axis when grown epitaxially on GaN. An applied strain along or perpendicular to the *c*-axis will cause an internal displacement of the metal sublattice with respect to that of the nitrogen, effectively changing the polarisation of the material. This strain effect provides an additional contribution to the polarisation of the material, referred to as the piezoelectric component, and is particularly relevant to strained heterostructures.

Virtually all commercial GaN-based LEDs are grown along the *c*-axis of the crystal. Since this is a polar direction, there exists an electric field across the InGaN quantum well due to a difference in polarisation for the well and barrier material. The electric field will cause a tilting of the conduction and valence bands in the well, separating the electrons and holes and shifting the quantum well emission wavelength to lower energy, as illustrated in Fig. 5.8. This is known as the quantum confined stark effect (QCSE).

There are some general observations about the QCSE relevant to nitride QWs: with the presence of an electric field, the transition energy is shifted to a lower value (from $\Delta E_{g,QW}$ to ΔE_{g1}) and this shift is roughly equal to the sum of the shifts of the first electron (ΔE_{e1}) and hole (ΔE_{h1}) levels; it is the hole state that contributes most due to the larger effective mass; electrons and holes are separated from each other spatially by the electric field across the quantum well, resulting in a reduced overlap of electron and hole wave functions and thus a longer radiative lifetime; wider wells (QW2) show more obvious effects of the QCSE and a larger potential drop (ΔE_{E2}) across the well. For a sufficiently wide well, the emission can be lower energy than the bandgap of the quantum well material itself.

The impact of the internal field, especially the piezoelectric field caused by strain, on quantum well recombination behaviour has been confirmed experimentally and reported in various III-nitride-based heterostructures [27–32]. Redshifts of emission energy and lower emission intensity were found in strained quantum wells based on III-nitrides, confirming the strong influence of the strain-induced piezoelectric field. However, with increasing carrier injection, a blue shift of the emission peak was observed by several researchers [33,34] and attributed to the reduction of the QCSE due to the in-well field screening by carriers. Therefore, in

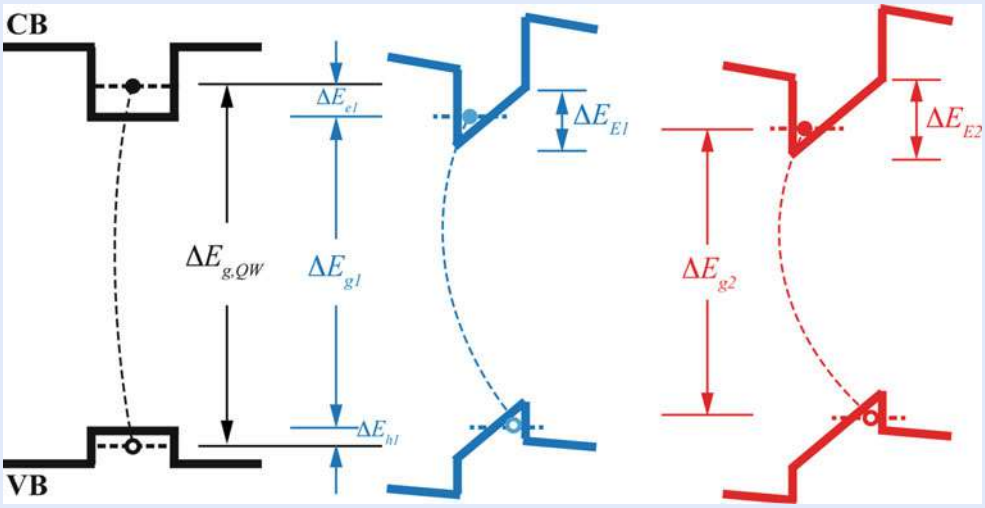


Fig. 5.8 Schematic plot showing the effects of the QCSE on InGaN/GaN quantum wells: *black*: QW1 without electric field; *blue*: QW1 with electric field; *red*: QW2 (thicker quantum well) with electric field

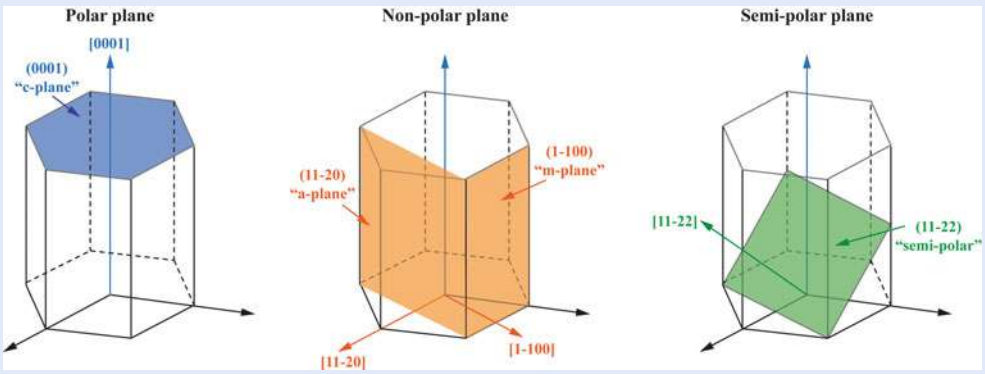


Fig. 5.9 Schematic of the principle polar, non-polar and semi-polar planes of GaN. The QCSE effect should be eliminated by growing along a non-polar direction such as [1–100] and [11–20] or minimised along a semi-polar direction such as [11–22]

an LED structure, the electric field across the quantum wells is not only determined by the polarisation field but also affected by the carrier density and distribution in the quantum well region. The carriers may be from carrier injection (optical or electrical), as well as from doping, either intentional dopants or non-intentional impurities.

From the discussion above, it is obvious that the QCSE is not desirable for LEDs of high efficiency and good colour consistency. Figure 5.9 shows the main polar, non-polar and semi-polar planes of GaN. In principle, the QCSE should be eliminated by growing along a non-polar direction such as [1–100] and [11–20] or minimised along a semi-polar direction such as [11–22]. The efficiency of non-polar and semi-polar light emitting structures is therefore expected to be enhanced over that of polar.

However, it was found that the defect density is currently much higher in GaN structures grown in such directions [35], unless expensive freestanding non-polar or semi-polar GaN substrates are used [36]. Furthermore, the indium incorporation in the InGaN MQWs grown along non-polar direction is 2–3

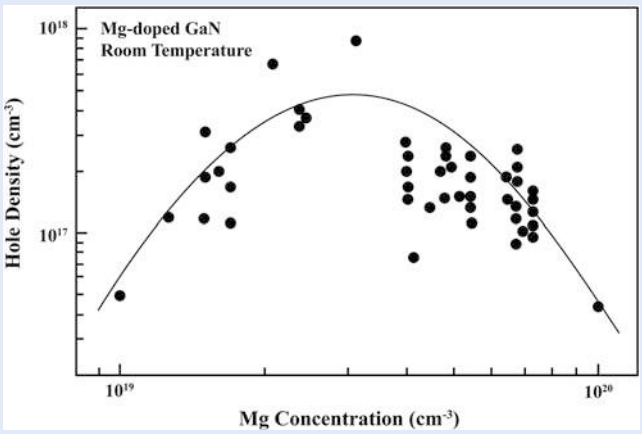
times lower than along the *c*-plane for similar growth conditions [37]. The output power of the non-polar LEDs also reduced dramatically when the emission wavelength was longer than 400 nm. Therefore, a non-polar plane is considered not suitable for LEDs with emission wavelengths longer than blue and semi-polar planes are preferred for blue, yellow and red LEDs with reduced internal field, but again high defect densities are a problem. Despite the potential advantages of reduced internal field, non-polar and semi-polar LEDs are currently not commercially viable due to their lower overall performance and the requirement of expensive freestanding GaN substrates.

5.5.3 *p*-Type Doping

For III-nitrides, *p*-type doping is problematic and the realisation of *p*-type conductivity was another major breakthrough in the historical development of nitride-based LEDs. Non-intentionally doped GaN usually shows *n*-type conductivity; however, the improvement in crystal growth methods has managed to reduce this background doping level sufficiently to allow controllable *p*-type doping [38]. Many potential *p*-type dopants have been tried and so far magnesium is the most successful *p*-type dopant for GaN, AlGaN and InGaN with low Al and In mole fractions.

There are two main issues involved in Mg doping: (1) the presence of hydrogen in MOVPE and HVPE growth environments results in the passivation of Mg by forming Mg–H complexes that are electrically inactive; (2) Mg forms relatively deep acceptor states ~160–200 meV above the valence band [39], resulting in only a small fraction activated at room temperature and therefore low conductivity of *p*-type GaN. This means the hole concentration will always be more than an order of magnitude lower than the Mg concentration. Furthermore, heavily Mg-doped GaN is subject to self-compensation due to the formation of donor-like structural defects [40].

The first issue can be solved by thermal annealing under an N_2 ambient at a temperature higher than 700 °C [41,42] or by electron beam irradiation [43] to activate the passivated Mg. The thermal annealing technique has become the standard method for dopant activation because it is straightforward, reliable and can be implemented in-situ, within the MOVPE growth reactor. In contrast, the second issue of a deep acceptor level and self-compensation is intrinsic and is the main reason limiting the hole concentration. ◘ Figure 5.10 shows the



◘ **Fig. 5.10** Hole density in GaN:Mg films as determined by the Hall effect versus the Mg concentration of the films as measured by SIMS. Data from Obloh et al. [44]

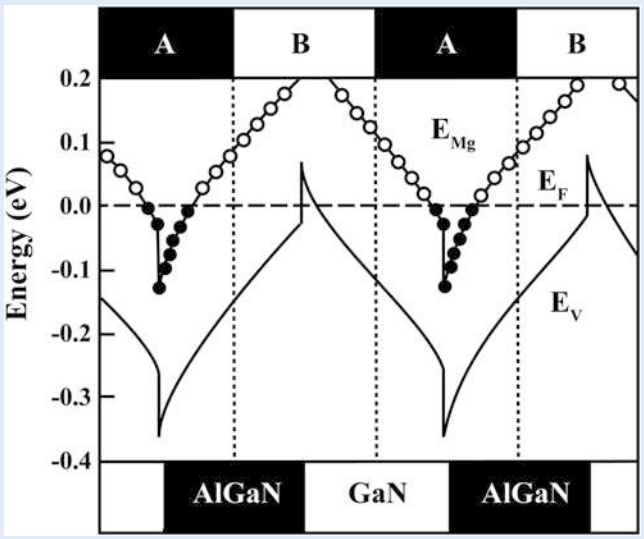


Fig. 5.11 Calculated valence band diagram for the Mg-doped $Al_{0.2}Ga_{0.8}N$/GaN superlattice with spontaneous and piezoelectric polarisation fields taken into account. The *dashed line* indicates the Fermi energy and the *circles* represent the energy of the Mg acceptor with *solid circles* indicating the ionised form. The growth direction for normal Ga-polarity material would be from left to right. Data from Kozodoy et al. [46]

concentration of free holes at room temperature in Mg-doped GaN by MOVPE [44]. The hole concentration reaches its maximum value of about 10^{18} cm^{-3} for a Mg concentration of about 3×10^{19} cm^{-3}, and thereafter decreases with further increase of Mg doping.

A promising method to achieve higher acceptor activation and lower electrical resistivity is to use AlGaN/GaN superlattices. This provides a periodic oscillation in the valence band edge, allowing ionisation of acceptors in the wide bandgap AlGaN layers to provide hole accumulation in the adjacent GaN layers, leading to an overall increase in hole concentration [45]. The principle is illustrated in Fig. 5.11, where it is apparent that polarisation fields in the nitrides enhance the band edge modulation, leading to parallel sheets of highly concentrated free carriers where the Fermi level intersects the valence band [46]. This can result in spatially averaged hole concentrations in the 10^{18} cm^{-3} range for such superlattices [47,48]. Using the same approach, *p*-type conductivity in $Al_{0.17}Ga_{0.83}N/Al_{0.36}Ga_{0.64}N$ superlattices has been demonstrated [49], and this will undoubtedly be a common approach in deep-UV emitting LEDs where *p*-type AlGaN is even more problematic due to wider bandgaps.

Although the development of *p*-type doping has enabled high-efficiency semiconductor devices, the hole carrier concentration in a GaN-based LED device is still about two orders of magnitude lower than the electron concentration, resulting in a large asymmetric carrier distribution in the active region. *P*-type doping in GaN and its alloys with InN and AlN remains a topic of interest at both a fundamental science level and in technological aspects.

5.5.4 Green Gap and Efficiency Droop

In spite of the challenges mentioned above, the performance of nitride LEDs has continued to advance, and devices emitting in the violet to green spectral region have already been commercialised. The highest efficiencies are still achieved for

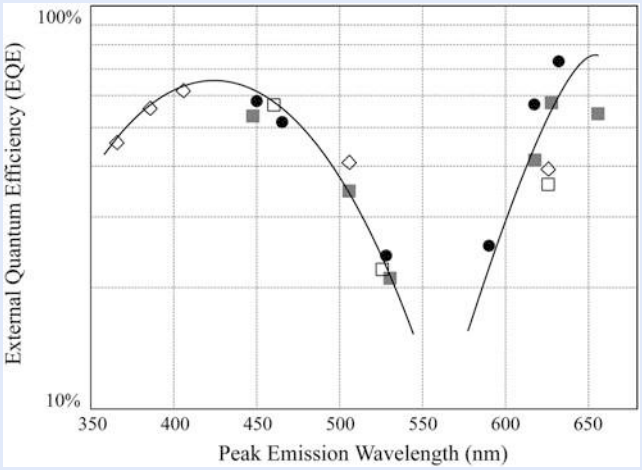


Fig. 5.12 Plot of the external quantum efficiency (EQE) of commercial LED devices measured using EL at 350 mA, showing the issue of the 'green gap'

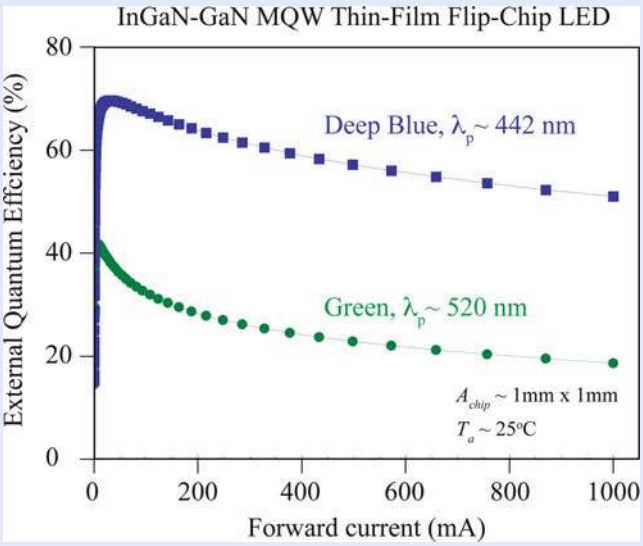


Fig. 5.13 Plot of the external quantum efficiency (EQE) and light output of commercial blue and green emitting LED devices measured using EL at different forward current, showing the issues of 'efficiency droop' at higher current density

blue and violet wavelengths, and despite considerable research efforts (both academic and industrial), a rapid drop in performance towards deep green (the 'green gap') and UV wavelengths remains (Fig. 5.12). Another important problem is that the efficiency of InGaN-based LEDs decreases with increasing current density, an effect known as 'efficiency droop' (Fig. 5.13). Solving the 'green gap' and 'efficiency droop' problems is currently a key focus for research both in academia and industry [12,50–56].

For AlGaInP LEDs, the reason for the lower efficiency at wavelengths shorter than 600 nm is the transition from a direct to an indirect bandgap, as shown in Fig. 5.1. The factors limiting the IQE of nitride LEDs are complex and not well understood. For InGaN, the reason for decreased efficiency in the green spectral region has been attributed to the miscibility gap between GaN and InN [57] and

high polarisation fields caused by the increasing strain with higher InN mole fractions.

Possible mechanisms of 'efficiency droop' that have been proposed include Auger recombination [52,56], high defect density [54,58], carrier leakage [59], polarisation-induced built-in electric fields at hetero-interfaces [60,61], poor *p*-type conductivity [62,63] and carrier delocalisation at high current densities [64]. In order to reduce the current density and thus the efficiency droop, a thicker single quantum well has been proposed to replace thin multiple quantum wells as the active region [12]. However, it was found that thicker InGaN QWs are only feasible in the short wavelength range around 400 nm. For LEDs emitting at longer emission wavelengths, the material quality decreases due to growth at lower temperatures and the internal field rapidly increases due to higher In contents. Therefore, most commercial blue- and green-emitting LEDs still use thin multiple quantum wells as the active region.

5.5.5 Chip Design

The discussions above on crystal growth, *p*-type doping, internal fields and efficiency droop are mainly concerned with how to improve the internal efficiency of GaN-based LEDs by optimising the material growth and the structure design. However, improving the generation of light in the active region alone is not enough to achieve an efficient LED device, because the overall efficiency of an LED device is determined by many components as mentioned earlier in this chapter. Chip design is an important area of research to reduce internal reflection for higher light extraction and to enable uniform current injection (especially hole injection). The schematic structures of several different chip designs developed over the years are illustrated in Fig. 5.14.

Compared with the *n*-type region, the *p*-type layer is very resistive and of limited thickness. To overcome the current spreading problem, a semi-transparent NiAu contact was originally deposited over the *p*-GaN [66] for a conventional shape LED chip. However, this approach results in significant losses when the emitted light passes through the *p*-contact. The 'Flip-chip' (FC) approach was then developed, where the LED chip is inverted and the light is emitted from the *n*-GaN side. In this approach, the NiAu contact is replaced by a thick and reflective contact, usually comprising silver, to reflect back the light emitted towards the *p*-type layer side [67]. In order to overcome the internal reflection problem, laser lift-off of the sapphire substrate and *n*-GaN roughening were used in the thin-film flip-chip (TFFC) LED design, achieving light extraction efficiency as high as 80 % by 2006 [68]. A similar vertical thin-film device (VTF) was also developed, resulting in an estimated light extraction efficiency of 75 % [69]. In recent years, patterned sapphire substrates have become very popular due to the advantages of improved material quality and ease of light extraction. Combining patterned sapphire substrates with an indium-tin-oxide (ITO) current spreading layer, a light extraction efficiency as high as 88 % was estimated [70] for this PSS-ITO approach.

The above approaches all extract the emitted light primarily from the top or bottom side of the LED chip. When a bulk GaN substrate is used, the sidewalls of the LEDs can be used to extract part of the light through geometric die shaping, as shown in Fig. 5.15. These volumetric LEDs have the potential to achieve even higher light extraction efficiency than thin-film LEDs based on modelling [71]. Today, light extraction efficiencies exceeding 85 % are achieved for high power TFFC InGaN LEDs [12]. When using GaN as a substrate, the light extraction efficiency can also be as high as 90 %.

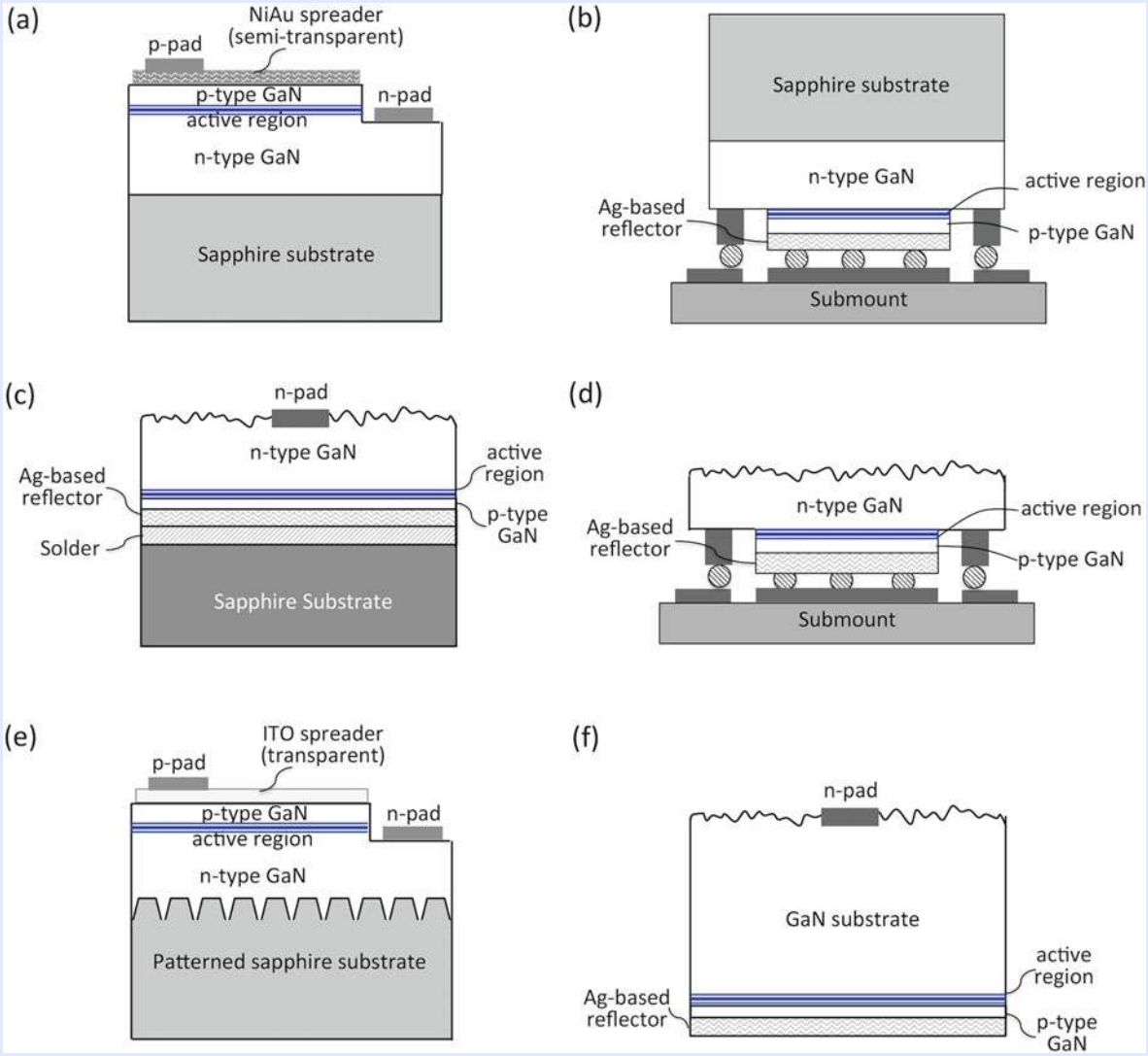


Fig. 5.14 Schematic cross sections of various GaN-based LED chip designs: (**a**) Conventional chip. (**b**) Flip-chip (FC). (**c**) Vertical thin film (VTF). (**d**) Thin-film flip-chip (TFFC). (**e**) Patterned sapphire substrate combined with ITO contact (PSS-ITO). (**f**) GaN substrate volumetric LED chips. After Nakamura and Krames [65]

5.5.6 Generation of White Light with LEDs

Whereas LEDs emit light of a single colour in a narrow wavelength band, white light is required for a huge range of applications, including LED backlighting for large LCD displays, and general home and office lighting. White light is a mixture of many colours (wavelengths) and there are two main methods to generate white light using LEDs: Phosphor method and RGB method, as illustrated in Fig. 5.16.

The first commercially available white LED was based on an InGaN chip emitting blue light at a wavelength of 460 nm that was coated with a cerium-doped yttrium aluminium garnet (YAG) phosphor layer that converted some of the blue light into yellow light [72]. Nearly all white LEDs sold today use this method. The phosphor layer is sufficiently thin that some blue light is transmitted through it, and the combination of blue and yellow produces a 'cool white' light.

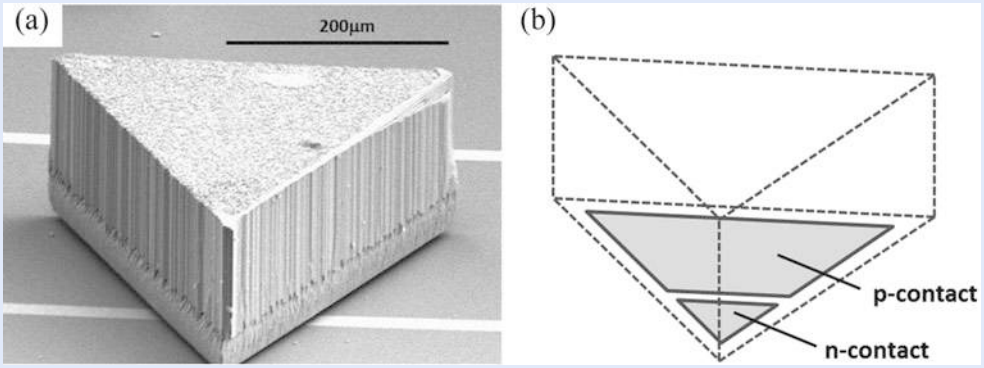


Fig. 5.15 (**a**) Scanning electron microscope image of a fabricated triangular-shaped gallium nitride on gallium nitride (GaN-on-GaN) LED chip with roughened top and side surfaces. (**b**) Corresponding device geometry. Unlike cuboidal shapes, light is not trapped inside by total internal reflection with this geometry. Reprinted with permission from David et al. [71]. Copyright 2014, AIP

Fig. 5.16 Illustration of white light generation using phosphor method (blue/UV LEDs + phosphors) and using RGB method (Red + Green + Blue LEDs)

This is fine for many applications (displays, lighting in cars, buses, yachts and cell phones back lights), but the quality of light is probably not good enough for home lighting, for which a warmer white light containing some red light is desirable. To generate 'warm white', red phosphors are typically added [73].

Since the efficiency with which existing red phosphors are excited using blue light is much less than that using near-UV light, a better route to generate 'warm white' light might be to use a near-UV LED plus red, green and blue or more coloured phosphors. Thick phosphor layers would be used so that no near-UV light from the LED would be transmitted in much the same way as the phosphor coating on fluorescent tubes and CFLs prevents the transmission of UV light. The drawback of this method is the large intrinsic energy loss from converting a near-UV photon to a lower energy visible photon.

Mixing red and green and blue (RGB) LEDs is an alternative way to produce white light without using phosphors, which is potentially the most efficient. However, there are three basic problems with this method. The first is that the efficiency of green LEDs is much less than that of red and blue, for reasons that are not yet understood (this is known as the 'green gap' problem described earlier). Hence the overall efficiency of this method is limited by the low efficiency of the green. Second, the efficiencies of red, green and blue LEDs change over time at different rates. Hence if a high quality white light is produced initially, over time the quality of the white light could degrade noticeably. However, this process is

slow and can be corrected electrically using automatic feedback. Third, because the emission peaks of LEDs are narrower than those of most phosphors, red plus green plus blue LEDs will give a poorer colour rendering than by using phosphors. This problem can be minimised by a careful choice of LED emission wavelengths, and of course, more than three different colour LEDs can be used for better coverage of the visible spectrum. In particular, using four LEDs (red, yellow, green and blue) can give a good colour rendering, although at the expense of increased complexity.

5.5.7 LED Packaging

LED packaging secures and protects the LED chips from damage caused by electrostatic discharge, moisture, high temperature and chemical oxidation. When designing the LED package, the issues involved in optical control, thermal management, reliability and cost need to be addressed simultaneously. The main package components include the LED die/chip, electrodes (anode and cathode), bond wire (connecting the LED die and electrodes), heat sink (removing heat generated by the LED die), phosphor coating (for white light emission) and primary lens (for directing the light beam).

Many solutions have been developed over the years for high power LED packages, as shown in Fig. 5.17, ranging from single large die packaging with input powers of 1–2 W to chip-on-board and 'Jumbo Die' solutions that can take input powers up to 94 W with lumen flux higher than 10,000 lm from a single package. Depending on the application, different LED package sizes and powers would be required. An interesting trend of LED packaging is to move from chip-based packaging to wafer-level packaging, with advantages of higher packing density, ease of integration on circuit boards, higher current density and higher reliability.

Fig. 5.17 Wide variety of solutions for high power LED packages. Images from Philips Lumileds, Osram, Cree and Luminus

5.6 LEDs for Lighting

Over the last decade, advances in the material quality, LED structure, chip architecture and package design have improved the performance of LEDs dramatically in terms of light quality, efficiency/efficacy, lifetime and cost. This has enabled LEDs to become a realistic replacement of traditional light sources such as incandescent and fluorescent lamps.

5.6.1 Quality of LED Lighting

People have become used to high quality lighting provided by conventional light sources, especially those installed at home, such as incandescent and halogen lamps. Colour temperature, colour rendering index (CRI) and colour consistency are the main factors when evaluating the quality of a white light source.

The planckian black-body radiation spectrum is used as a standard for white light because its spectrum can be described using only one parameter, namely the colour temperature. The colour temperature (CT) or correlated colour temperature (CCT) of a white light source, given in units of Kelvin, is defined as the temperature of a planckian black-body radiator whose colour is closest to that of the white light source. With increasing temperatures, a planckian black-body radiator glows in the red, orange, yellowish white, white and ultimately bluish white. Therefore, the colour temperate of a white light source can be used to describe its appearance. For conventional lighting technologies, the CCT spans a wide range, from 2700 to 6500 K. 'Warm white' light, such as from incandescent lamps, has a lower colour temperature (2700–3500 K), while 'cool white', which is a more blue–white, has a higher colour temperature (3500–5500 K). 'Warm white' is in the most common lamp colour used in residential lighting in the USA and Europe.

Another important characteristic of a white light source concerns how precisely the different colours of an object show up under illumination from the light source. This is measured in terms of the CRI. Some examples of different light sources and their corresponding spectrum are shown in Fig. 5.18. An ideal light source, such as sunlight, can reproduce colours perfectly and has a CRI of 100. Natural light LED lamps or full spectrum LED lamps, e.g. white LEDs based on near-UV LEDs plus RGB phosphors technology [107] have a CRI value as high as 95. Therefore the colours under full spectrum LED lamps also appear to be rich and vivid, similar to those under sunlight. 'Warm white' LEDs usually have a CRI higher than 80, which is acceptable to replace conventional light sources for most cases. While for conventional 'cool white' LEDs, the colour reproduction becomes insufficient, similar to fluorescent light.

It is noted that the current CIE colour rendition system of eight test colours to determine the CRI of a light source is designed around conventional light source technology and is not sufficient for LEDs. A new and better method of measuring and rating colour rendition for LED light sources is under development. For lighting professionals, the specific spectrum of a particular light source or the position of the colour points of a light source in relation to the black-body locus is a more accurate way of determining the value of the colour rendition.

Conventional light sources, such as incandescent and halogen lamps, have good colour consistency during their lifespan. For LED lighting, achieving good colour consistency is challenging. The colour distribution of blue LEDs and phosphors may result in greenish, blueish and pinkish white light. Furthermore, the colour of LEDs can shift with temperature and time. LED manufactures have put a lot of efforts into understanding and controlling the colour shift of LEDs.

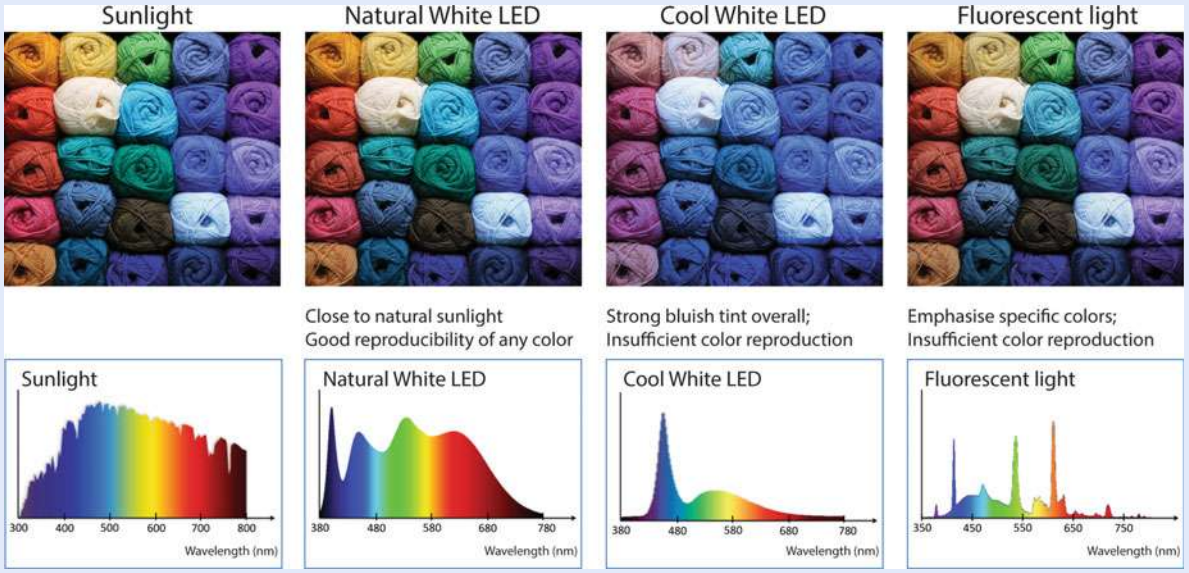


Fig. 5.18 Examples of different light sources and their corresponding spectrum. A broader spectral source more accurately renders colours of illuminated objects Image from online resources

The uniformity of epitaxy, processing and phosphor technologies are improving continuously, enabling a tighter distribution of LEDs in the production process. The LED industry has also adopted a strict binning system to ensure colour consistency between LEDs. Meanwhile, LED industry standards and regulations are being developed. For example, in EU directive (EU-1194/2012), one of the functionality requirements is on colour consistency and a variation of chromaticity coordinates within a six-step MacAdam ellipse or less is required [87]. Some manufactures have implemented LED lighting products that fall within a single three-step MacAdam ellipse to avoid a difference in colour between two sources that may be perceived [74, 90].

Since LEDs have different colours at different temperatures, leading LED manufactures now specify their LEDs at real application temperatures (85 °C), instead of a 25 °C operating temperature, on their datasheet to ensure the customers receive the exact colour intended. Although the colour consistency of LED lighting has improved greatly, the colour shift during its long lifetime remains a large area of concern. The solutions rely on a better understanding of the degradation mechanisms of LED chips and other components with time. Considering the rapid improvement made during the short LED history so far, we have every reason to believe that within a short time LED lighting technology will totally surpass conventional light sources in both quantity and quality.

5.6.2 Efficacy

Radiometric units, such as optical power in watts (W), are used to characterise light in terms of physical quantities. However, the human eye is sensitive only to light in the visible spectrum, ranging from violet (with a wavelength of ~400 nm) through to red (with a wavelength of ~700 nm) and has different sensitivity at different wavelengths, as shown in Fig. 5.19. The maximum sensitivity of the human eye is to green light with a wavelength of 555 nm. Therefore, to represent the light output of an optical source as perceived by the human eye, photometric

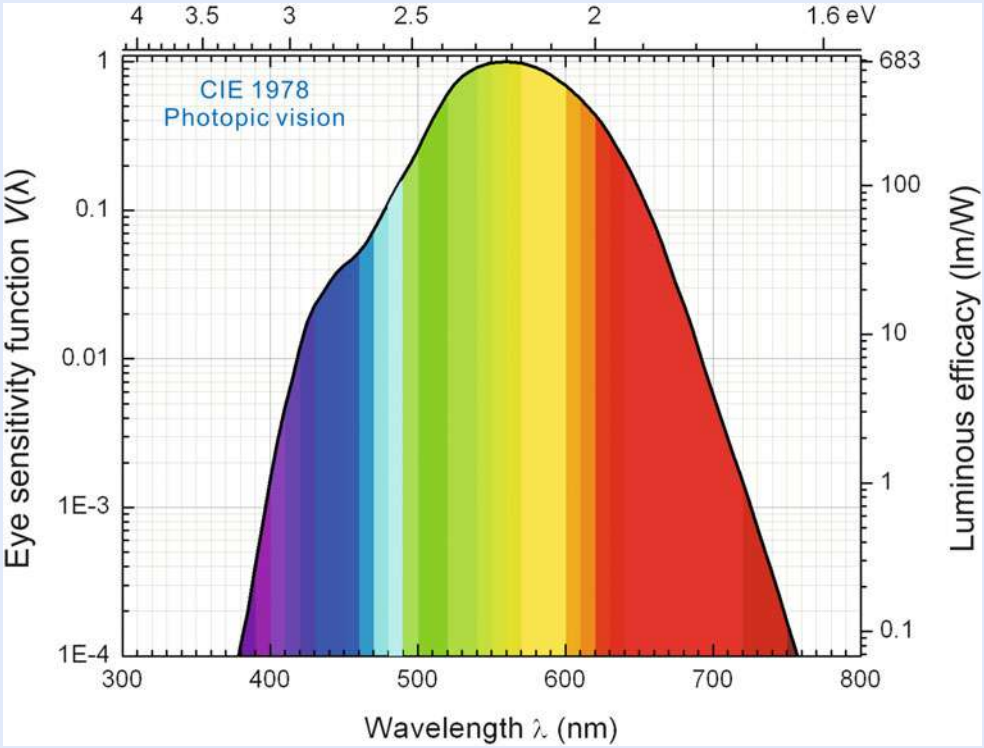


Fig. 5.19 Eye sensitivity function V(λ) and luminous efficacy in lm/W. The maximum sensitivity of the human eye is to green light with a wavelength of 555 nm (Data after 1978 CIE) [81]. It is noted that definition of the luminous efficacy here is the light power output in lumen divided by the optical power in (measured in W), rather than electrical power in

units, such as lumens (lm), are used instead of radiometric units. The efficacy of a light source takes into account the sensitivity of human vision, so that green light contributes more strongly to efficacy than blue or red light, and ultraviolet and infrared wavelengths do not contribute at all. The unit of efficacy is lumens per watt (lm/W), corresponding to light output power (as perceived by the human eye and measured in lumens) relative to electrical input power (measured in Watts).

It should be noted that there is a fundamental trade-off between efficacy and colour rendering [75]. The corresponding colour temperature should also be considered when comparing the efficacy of different white light sources. Generally speaking, a 'warm white' LED source of high CRI usually has lower efficacy compared with a 'cool white' LED source of lower CRI. The highest reported efficacy so far from a packaged LED device is 303 lm/W at a drive current of 350 mA and with a correlated colour temperature of 5150 K [76].

5.6.3 Lifetime

One of the main advantages of LED lighting is its long lifetime that potentially can span to 50,000 h or even 100,000 h. Similar to all electric light sources, LED lighting experiences a decrease in the amount of light emitted over time, a process known as lumen depreciation. For general lighting purpose, the useful life of an LED is defined as the point at which light output has declined to 70 % of initial lumens. The primary cause of LED lumen depreciation is heat generated at the LED junction that will affect the performance of key LED package components as

well as materials [77]. Heat management is therefore an important factor in determining the effective useful life of the LED. The lifespan of commercial LED replacement lamps is already longer than 15,000 h (some are longer than 25,000 h). As LEDs become more efficient over time, the problem of heat management will largely disappear and a longer lifetime of LED lighting is expected. The lifetime of LED lamps is also limited by the shorter lifetime of the control electronics used. So more attention is being paid to the development of sophisticated control electronics for LED lighting.

5.6.4 Cost

Cost is probably the major factor limiting the widespread use of white LEDs in our homes and offices. GaN-based LED replacement lamps are significantly more expensive than filament light bulbs or compact fluorescent lamps (CFLs). However, the cost per lumen is continuously decreasing, following the Haltz's law (see Fig. 5.20).

It should be noted that the total ownership cost of lighting includes energy savings and replacement cost, which makes LEDs more competitive, compared to conventional lighting technologies. Nevertheless, in order to achieve significant market penetration, the initial cost ($/klm) of LEDs needs to be reduced 10 times to be comparable to the cost of CFLs. To achieve the required cost reduction, many aspects of the manufacturing process will need to be addressed in parallel, as illustrated in Fig. 5.21. This diagram shows that the cost reduction shouldn't be based on sacrificing the three main LED quality factors: efficiency, reliability and customer experience. To make sure LED lighting remains a high quality light source, many aspects including LED materials, chip design, white light generation, component design, power supply circuit, luminaire optical and thermal design need to be taken care of.

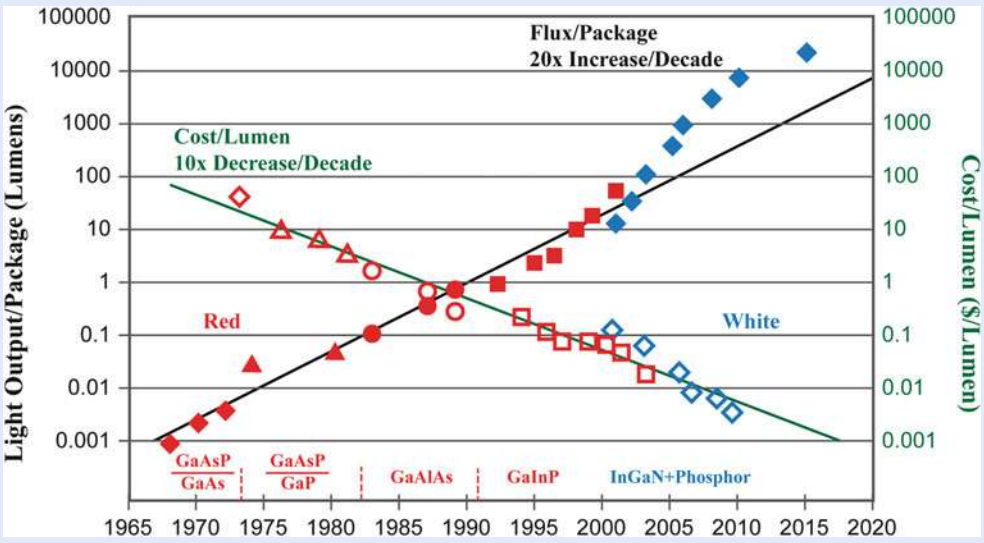


Fig. 5.20 Haltz's law showing that every decade, the cost per lumen falls by a factor of 10, and the amount of light generated per LED package increases by a factor of 20

Fig. 5.21 Integrated systems approach to solid-state-lighting manufacturing (after Mark McClear, Cree, Inc., SSL manufacturing workshop, Vancouver, OR, June 2009)

5.7 LED Lighting Applications: The Present and Future

Although significant improvements are still expected, the present performance of nitride LEDs is nevertheless superior in many respects compared with conventional lighting. LEDs are compact, efficient, long-lasting and controllable, and are already widely used, for example (as shown in Fig. 5.22), as traffic signals, in large outdoor displays, as interior and exterior lighting in aircraft, cars and buses, as bulbs in flash lights and as backlighting for LCD TVs cell phones and displays. Due to their long lifetime, LEDs are also being fitted on airport runways, where the operational cost can be significantly lowered: traditional lighting on runways lasts for about 6 months, and the runway has to be closed to replace it, at considerable cost. The performance of LEDs improves at lower temperatures, which is perfect for illuminating refrigerated displays in supermarkets, where CFLs give poor performance because their efficiency is very low when cold. Architectural lighting also favours LEDs, which combine art due to the flexibility in use of LEDs, with energy saving and eco-friendliness.

The research and applications of LED lighting in the horticultural industry (Fig. 5.23) have also attracted a lot of attention [108], with benefits including better control of plant growth, increased yield, earlier flowering, faster root growth and more economical use of space. The lower electricity consumption and controllable light spectrum design are especially attractive features of LED lighting for horticulture applications.

Optogenetics is a new area of neuroscience that uses light to stimulate targeted neural pathways in the brain to uncover how neurons communicate and give rise to more complex brain functions. One key technical challenge in optogenetics is the realisation of a reliable implantable tool to precisely deliver light to the targeted neurons and to simultaneously record the electrical signals from the individual neuron. Such a neural probe requires the successful integration of light sources, detectors, sensors and other components on to an ultrathin cellular-scale injection needle, which can be inserted deep into the brain with minimum damage of tissue. Micro-LEDs are an ideal light source for this application due to the small size and controllable emission wavelength.

Fig. 5.22 Various application examples of LEDs as retail light, as backlighting for LCD TVs, as outdoor street light, as bicycle light, as exterior lighting in cars, as architectural lighting and as airport runway light. Images from Osram, with permission

Fig. 5.23 Philips LED lighting for fast-track growth in horticulture. Image from Philips, with permission

Visible light communication (VLC) technology, more recently referred to as Li-Fi (Light Fidelity), transmits data using light sources that modulate intensity faster than the human eye can perceive. Although still in its infancy, VLC is believed to be a future technology in wireless communication. LEDs are especially suitable for this application due to their fast switch on/off rate and long lifetime. By using an array of micro-LEDs, instead of conventional LEDs, the data transmit rate can be increased to more than 10 Gbps (Gigabits per second). An even bigger picture of this technology is to combine information displays, lighting and

Table 5.2 Comparison of LED light bulbs with conventional classic light bulbs

	Incandescent	Halogen incandescent	Compact florescent	LED light bulbs
				
Lumen	1100	1200	970	1055
Power (W)	75	70	15	13
Efficacy (lm/W)	15	17	65	81
Colour temperature (K)	2700	2800	2700	2700
Colour rendering index	100	100	81	80
Rated lifetime (h)	750–2000	2000	10,000	15,000
Mercury content (mg)	0	0	≤2	0
Warm-up time to 60 % light	Instant full light	Instant full light	5–40 s	Instant full light
Sales price	Banned [109]	£2.00	£5.00	£10.00

high-bandwidth communications in a single system, which will bring revolutionary solutions for machine-to-machine communications, smart homes and vehicles, mobile communications, imaging systems, personal security, healthcare and so on.

5.7.1 General Illumination and Energy Saving

Among all these exciting applications of LED lighting, those in general illumination, including residential, office, shop, hospitality, industrial, outdoor and architectural lighting, are the most relevant to our daily life and have the greatest energy saving potential. Both LED replacement classic light bulbs and LED fixtures are used for general illumination. A comparison of indoor LED light bulbs with other conventional light bulbs is given in Table 5.2, showing the advantages of LED lighting in energy saving without sacrificing performance. Due to its high initial cost, the current market penetration of LED lighting products is still very small. However, if the current trends in LED price and performance continue, LED lighting is projected to gain significant market penetration in USA, reaching 48 % of lumen-hour sales of the general illumination market by 2020, and 84 % by 2030.

Global population growth and urbanisation are increasing the overall demand for lighting products and the corresponding energy consumption by lighting. According to a recent US Department of Energy (DOE) report, lighting consumed ~18 % of total US electricity use in 2013, using approximately 609 TWh of electricity, or about 6.9 quads of source energy. LEDs are projected to reduce lighting energy consumption by 15 % in 2020 and by 40 % in 2030, saving 3.0 quads in 2030 alone. Assuming the current mix of generation power stations, these energy savings would reduce green house gas emission by approximately 180 million metric tons of carbon dioxide. Considering the global population growth, resource scarcity and climate change concerns, the development and adoption of LED technology is strategically important for a sustainable society.

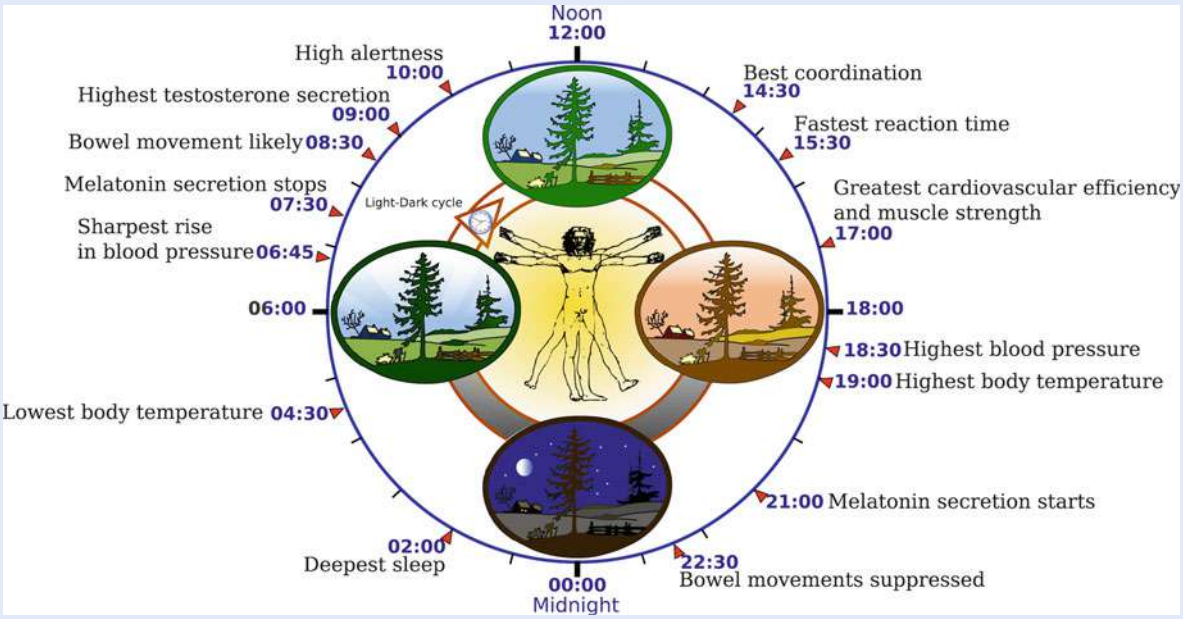


Fig. 5.24 Some features of the human circadian (24-h) biological clock. Image was done by Yassine Mrabet and uploaded by Addicted04 to Wikipedia, under creative CC BY-SA 3.0 free license

5.7.2 Circadian Rhythm Lighting

LED-based solid-state lighting is not just a replacement of traditional illuminations, but rather a multifunctional device we can use to improve our mood, health, productivity and much more. Because it is easily colour-tunable and dimmable, LED lighting is ideal to create circadian rhythm lighting that matches the needs of human biological cycles, or circadian rhythms, in the most effective and appropriate way.

Human beings are governed to some degree by an internal biological clock, called the circadian rhythm, as illustrated in Fig. 5.24. Light is the most powerful stimulus of the human body clock, and the timing of light exposure during the course of a day is responsible for how circadian rhythms are synchronised with the environment. For example, one of the best cures of 'jet lag' caused by circadian rhythm disruption is exposure to daylight to reset the body clock.

Modern industrialised society heavily relies on artificial lighting. Research tells us that circadian rhythm disruption through inappropriate artificial light causes many physical and mental health issues: fatigue, cancer, obesity, diabetes, depression, mood and sleep disorders, reduced physical and mental performance, reduced productivity and irritability are all related in some shape or form to a circadian system that isn't functioning properly. The most natural light is sunlight, which is dynamic and variable in brightness, colour temperature and spectral distribution during the day. Daylight provides bright blue-rich light in the early morning to deliver an alerting signal as we wake up and a warm, low-level light in the early evening to prepare our body to rest.

The dynamic features and spectral design flexibility of LED lighting enable the creation of personalised lighting to mitigate circadian rhythm disruption, optimise mood and visual experience, and improve our sense of wellbeing, in better ways than ever before. Combined with smart building control systems, LED circadian

rhythm lighting can be programmed to change colour temperature and light level automatically, allowing for the indoor reproduction of natural outdoor lighting conditions. Some circadian rhythm lighting products are already commercially available, for example, on aircraft for long-haul flights. In the future, we could expect LED lighting to become more intelligent and closer to natural light, contributing strongly to our health and wellbeing, as well as energy saving.

5.8 Chapter Summary

LED-based solid-state lighting promises to provide a high quality and energy efficient light source for our daily life. With continuous advances in efficiency and reductions in cost, LED lighting is on course to be the dominant form of lighting in homes, offices, cities and transport throughout the world. LED lighting is more than an energy efficient alternative to conventional light sources; it is suitable to create circadian rhythm lighting that can make us healthier and more productive. LED lighting is also intelligent and could interface with building management systems, transmit high-speed wireless data, fine-tune occupancy and functional sensing, and is an important integral part of our future smart home.

References

1. Holonyak N, Bevacqua SF (1962) Coherent (visible) light emission from $Ga(As_{1-x}P_x)$ junctions. Appl Phys Lett 1(4):82–83. doi:► 10.1063/1.1753706
2. Nakamura S, Senoh M, Mukai T (1993) High-power InGaN/GaN double-heterostructure violet light emitting diodes. Appl Phys Lett 62:2390–2392. doi:► 10.1063/1.109374
3. Nakamura S, Pearton S, Fasol G (1997) The blue laser diode. Springer, Berlin
4. Vurgaftman I, Meyer JR, Ram-Mohan LR (2001) Band parameters for III-V compound semiconductors and their alloys. J Appl Phys 89:5815–5875. doi:► 10.1063/1.1368156
5. Vurgaftman I, Meyer JR (2003) Band parameters for nitrogen-containing semiconductors. J Appl Phys 94:3675–3696. doi:► 10.1063/1.1600519
6. Saitoh T, Kumagai M, Wang H, Tawara T, Nishida T, Akasaka T, Kobayashi N (2003) Highly reflective distributed Bragg reflectors using a deeply semiconductor/air grating for InGaN/GaN laser diodes. Appl Phys Lett 82:4426–4428. doi:► 10.1063/1.1586992
7. Wu J, Walukiewicz W, Yu KM, Shan W, Ager JW III, Haller EE, Lu H, Schaff WJ, Metzger WK, Kurtz S (2003) Superior radiation resistance of In1-xGaxN alloys: full-solar-spectrum photovoltaic materials system. J Appl Phys 94:6477–6482. doi:► 10.1063/1.1618353
8. Munoz E, Monroy E, Pau JL, Calle F, Omnes F, Gibart P (2001) III nitrides and UV detection. J Phys: Condens Matter 13:7115–7137. doi:► 10.1088/0953-8984/13/32/316
9. Xing H, Keller S, Wu YF, McCarthy L, Smochkova IP, Buttari D, Coffie R, Green DS, Parish G, Heikman S, Shen L, Zhang N, Xu JJ, Keller BP, DenBaars SP, Mishra UK (2001) Gallium

nitride based transistors. J Phys: Condens Matter 13:7139–7157. doi:► 10.1088/0953-8984/13/32/317
10. Mishra UK, Parikh P, Wu YF (2002) AlGaN/GaN HEMTs—an overview of device operation and applications. Proc IEEE 90:1022–1031. doi:► 10.1109/JPROC.2002.1021567
11. Schubert EF (2006) Light-emitting diodes, 2nd edn. Cambridge University Press, Cambridge
12. Laubsch A, Sabathil M, Baur J, Peter M, Hahn B (2010) High-power and high-efficiency InGaN-based light emitters. IEEE Trans Electron Devices 57:79–87. doi:► 10.1109/TED.2009.2035538
13. Popovici G, Morkoç H, Noor Mohammed S (1998) Deposition and properties of group III nitrides by molecular beam epitaxy. In: Gil B (ed) Group III nitride semiconductor compounds: physics and applications. Oxford University Press, Oxford, p 19
14. Porowski S, Grzegory I (1997) Growth of GaN single crystals under high nitrogen pressure. In: Pearton SJ (ed) GaN and related materials. Overseas, Amsterdam, p 295
15. Koleske DD, Coltrin ME, Cross KC, Mitchell CC, Allerman AA (2004) Understanding GaN nucleation layer evolution on sapphire. J Cryst Growth 273:86–99. doi:► 10.1016/j.jcrysgro.2004.08.126
16. Figge S, Bottcher T, Einfeldt S, Hommel D (2000) In situ and ex situ evaluation of the film coalescence for GaN growth on GaN nucleation layers. J Cryst Growth 221:262–266. doi:► 10.1016/S0022-0248(00)00696-5
17. Kappers MJ, Datta R, Oliver RA, Rayment FDG, Vickers ME, Humphreys CJ (2007) Threading dislocation reduction in (0001) GaN thin films using SiN_x interlayers. J Cryst Growth 300:70–74. doi:► 10.1016/j.jcrysgro.2006.10.205
18. Kappers MJ, Moram MA, Zhang Y, Vickers ME, Barber ZH, Humphreys CJ (2007) Interlayer methods for reducing the dislocation density in gallium nitride. Physica B 401–402:296–301. doi:► 10.1016/j.physb.2007.08.170
19. Cherns D, Henley SJ, Ponce FA (2001) Edge and screw dislocations as non-radiative centers in InGaN/GaN quantum well luminescence. Appl Phys Lett 78:2691–2693. doi:► 10.1063/1.1369610
20. Chichibu SF, Uedono A, Onuma T, Haskell BA, Chakraborty A, Koyama T, Fini PT, Keller S, DenBaars SP, Speck JS, Mishra UK, Nakamura S, Yamaguchi S, Kamiyama S, Amano H, Akasaki I, Han J, Sota T (2006) Origin of defect-insensitive emission probability in In-containing (Al, In, Ga)N alloy semiconductors. Nat Mater 10:810–816. doi:► 10.1038/nmat1726
21. Graham DM, Soltani-Vala A, Dawson P, Smeeton TM, Barnard JS, Kappers MJ, Humphreys CJ, Thrush EJ (2005) Optical and microstructural studies of InGaN/GaN single-quantum-well structures. J Appl Phys 97:103508. doi:► 10.1063/1.1897070
22. Hammersley S, Badcock TJ, Watson-Parris D, Godfrey MJ, Dawson P, Kappers MJ, Humphreys CJ (2011) Study of efficiency droop and carrier localization in an InGaN/GaN quantum well structure. Phys Status Solidi C 8:2194–2196. doi:► 10.1002/pssc.201001001
23. Humphreys CJ (2007) Does in form in-rich clusters in InGaN quantum wells? Philos Mag 87:1971–1982. doi:► 10.1080/14786430701342172
24. Oliver RA, Bennett SE, Zhu T, Beesley DJ, Kappers MJ, Saxey DW, Cerezo A, Humphreys CJ (2010) Microstructural origins of localisation in InGaN quantum wells. J Phys D Appl Phys 43:354003. doi:► 10.1088/0022-3727/43/35/354003
25. Smeeton TM, Kappers MJ, Barnard JS, Vickers ME, Humphreys CJ (2003) Electron-beam-induced strain within InGaN quantum wells: False indium "cluster" detection in the transmission electron microscope. Appl Phys Lett 83:5419–5421. doi:► 10.1063/1.1636534
26. Watson-Parris D, Godfrey MJ, Oliver RA, Dawson P, Galtrey MJ, Kappers MJ, Humphreys CJ (2010) Energy landscape and carrier wave-functions in InGaN/GaN quantum wells. Phys Status Solidi C 7:2255–2258. doi:► 10.1002/pssc.200983516
27. Aumer ME, LeBoeuf SF, Bedair SM, Smith M, Lin JY, Jiang HX (2000) Effects of tensile and compressive strain on the luminescence properties of AlInGaN/InGaN quantum well structures. Appl Phys Lett 77:821–823. doi:► 10.1063/1.1306648
28. Aumer ME, LeBoeuf SF, Moody BF, Bedair SM, Nam K, Lin JY, Jiang HX (2002) Effects of tensile, compressive, and zero strain on localized states in AlInGaN/InGaN quantum-well structures. Appl Phys Lett 80:3099–3101. doi:► 10.1063/1.1469219
29. Aumer ME, LeBoeuf SF, Moody BF, Bedair SM (2001) Strain-induced piezoelectric field effects on light emission energy and intensity from AlInGaN/InGaN quantum wells. Appl Phys Lett 79:3803–3805. doi:► 10.1063/1.1418453
30. Leroux M, Grandjean N, Massies J, Gil B, Lefebvre P, Bigenwald P (1999) Barrier-width dependence of group-III nitrides quantum-well transition energies. Phys Rev B 60:1496–1499. doi:► 10.1103/PhysRevB.60.1496

31. McAleese C, Costa PMFJ, Graham DM, Xiu H, Barnard JS, Kappers MJ, Dawson P, Godfrey MJ, Humpherys CJ (2006) Electric fields in AlGaN/GaN quantum well structures. Phys Status Solidi B 243:1551–1559. doi:► 10.1002/pssb.200565382
32. Wetzel C, Takeuchi T, Amano H, Akasaki I (1999) Piezoelectric Franz–Keldysh effect in strained GaInN/GaN heterostructures. J Appl Phys 85:3786–3791. doi:► 10.1063/1.369749
33. Kuroda T, Tackeuchi A (2002) Influence of free carrier screening on the luminescence energy shift and carrier lifetime of InGaN quantum wells. J Appl Phys 92:3071–3074. doi:► 10.1063/1.1502186
34. Riblet P, Hirayama H, Kinoshita A, Hirata A, Sugano T, Aoyagi Y (1999) Determination of photoluminescence mechanism in InGaN quantum wells. Appl Phys Lett 75:2241–2243. doi:► 10.1063/1.124977
35. Johnston CF, Kappers MJ, Humphreys CJ (2009) Microstructural evolution of nonpolar (11–20) GaN grown on (1–102) sapphire using a 3D-2D method. J Appl Phys 105:073102. doi:► 10.1063/1.3103305
36. Zhong H, Tyagi A, Fellows N, Wu F, Chung RB, Saito M, Fujito K, Speck JS, DenBaars SP, Nakamura S (2007) High power and high efficiency blue light emitting diode on free-standing semipolar (10-1-1) bulk GaN substrate. Appl Phys Lett 90:233504. doi:► 10.1063/1.2746418
37. Yamada H, Iso K, Saito M, Masui H, Fujito K, DenBaars SP, Nakamura S (2008) Compositional dependence of nonpolar m-plane InxGa1-xN/GaN light emitting diodes. Appl Phys Express 1:041101. doi:► 10.1143/APEX.1.041101
38. Nakamura S, Senoh M, Mukai T (1991) Highly P-typed Mg-doped GaN films grown with GaN buffer layers. Jpn J Appl Phys 30:L1708–L1711. doi:► 10.1143/JJAP.30.L1708
39. Doverspike K, Pankove JI (1998) Doping in the III-nitrides. SEM SEMIMET 50:259–277
40. Kaufmann U, Kunzer M, Maier M, Obloh H, Ramakrishnan A, Santic B, Schlotter P (1998) Nature of the 2.8 eV photoluminescence band in Mg doped GaN. Appl Phys Lett 72:1326–1328. doi:► 10.1063/1.120983
41. Nakamura S, Iwasa N, Senoh M, Mukai T (1992) Hole compensation mechanism of P-type GaN films. Jpn J Appl Phys 31:1258–1266. doi:► 10.1143/JJAP.31.1258
42. Nakamura S, Mukai T, Senoh M, Isawa N (1992) Thermal annealing effects on P-type Mg-doped GaN films. Jpn J Appl Phys 31:L139–L142. doi:► 10.1143/JJAP.31.L139
43. Amano H, Kito M, Hiramatsu K, Akasaki I (1989) P-type conduction in Mg-doped GaN treated with low-energy electron beam irradiation (LEEBI). Jpn J Appl Phys 28:L2112–L2114. doi:► 10.1143/JJAP.28.L2112
44. Obloh H, Bachem KH, Kaufmann U, Kunzer M, Maier M, Ramakrishnan A, Schlotter P (1998) Self-compensation in Mg doped p-type GaN grown by MOCVD. J Cryst Growth 195:270–273. doi:► 10.1016/S0022-0248(98)00578-8
45. Schubert EF, Greishaber W, Goepfert ID (1996) Enhancement of deep acceptor activation in semiconductors by superlattice doping. Appl Phys Lett 69:3737–3739. doi:► 10.1063/1.117206
46. Kozodoy P, Hansen M, DenBaars SP, Mishra UK (1999) Enhanced Mg doping efficiency in Al0.2Ga0.8N/GaN superlattices. Appl Phys Lett 74:3681–3683. doi:► 10.1063/1.123220
47. Kozodoy P, Smorchkova YP, Hansen M, Xing H, DenBaars SP, Mishra UK, Saxler AW, Perrin R, Mitchel WC (1999) Polarization-enhanced Mg doping of AlGaN/GaN superlattices. Appl Phys Lett 75:2444–2446. doi:► 10.1063/1.125042
48. Yasan A, McClintock R, Darvish SR, Lin Z, Mi K, Kung P, Razeghi M (2002) Characteristics of high-quality p-type AlxGa1 − xN/GaN superlattices. Appl Phys Lett 80:2108–2110. doi:► 10.1063/1.1463708
49. Kim JK, Waldron EL, Li YL, Gessmann T, Schubert EF, Jang HW, Lee JL (2004) P-type conductivity in bulk AlxGa1 − xN and AlxGa1 − xN/AlyGa1 − yN superlattices with average Al mole fraction >20%. Appl Phys Lett 84:3310–3312. doi:► 10.1063/1.1728322
50. Cho J, Schubert EF, Kim JK (2013) Efficiency droop in light-emitting diodes: challenges and countermeasures. Laser Photonics Rev 7(3):408–421. doi:► 10.1002/lpor.201200025
51. Galler B, Lugauer HJ, Binder M, Hollweck R, Folwill Y, Nirschl A, Gomez-Iglesias A, Hahn B, Wagner J, Sabathil M (2013) Experimental determination of the dominant type of Auger recombination in InGaN quantum wells. Appl Phys Express 6:112101. doi:► 10.7567/APEX.6.112101
52. Kioupakis E, Rinke P, Delaney KT, Van de Walle CG (2011) Indirect Auger recombination as a cause of efficiency droop in nitride light-emitting diodes. Appl Phys Lett 98:161107. doi:► 10.1063/1.3570656
53. Meyaard DS, Lin GB, Cho J, Schubert EF, Shim H, Han SH, Kim MH, Sone C, Kim YS (2013) Identifying the cause of the efficiency droop in GaInN light-emitting diodes by correlating the onset of high injection with the onset of the efficiency droop. Appl Phys Lett 102:251114. doi:► 10.1063/1.4811558

54. Monemar B, Sernelius BE (2007) Defect related issues in the "current roll-off" in InGaN based light emitting diodes. Appl Phys Lett 91:181103. doi:► 10.1063/1.2801704
55. Rozhansky IV, Zakheim DA (2006) Analysis of the causes of the decrease in the electroluminescence efficiency of AlGaInN light-emitting-diode heterostructures at high pumping density. Semiconductors 40:839–845. doi:► 10.1134/S1063782606070190
56. Shen YC, Mueller GO, Watanabe S, Gardner NF, Munkholm A, Krames MR (2007) Auger recombination in InGaN measured by photoluminescence. Appl Phys Lett 91:141101. doi:► 10.1063/1.2785135
57. El-Masry NA, Piner EL, Liu SX, Bedair SM (1998) Phase separation in InGaN grown by metalorganic chemical vapor deposition. Appl Phys Lett 72:40–42. doi:► 10.1063/1.120639
58. Yang Y, Cao XA, Yan CH (2009) Rapid efficiency roll-off in high-quality green light-emitting diodes on freestanding GaN substrates. Appl Phys Lett 94:041117. doi:► 10.1063/1.3077017
59. Schubert MF, Xu JR, Kim JK, Schubert EF, Kim MH, Yoon SK, Lee SM, Sone CL, Sakong T, Park YJ (2008) Polarization-matched GaInN/AlGaInN multi-quantum-well light-emitting diodes with reduced efficiency droop. Appl Phys Lett 93:041102. doi:► 10.1063/1.2963029
60. Iso K, Yamada H, Hirasawa H, Fellows N, Saito M, Fujito K, DenBaars SP, Speck JS, Nakamura S (2007) High brightness blue InGaN/GaN light emitting diode on nonpolar m-plane bulk GaN substrate. Jpn J Appl Phys 46:L960–L962. doi:► 10.1143/JJAP.46.L960
61. Xu JR, Schubert MF, Noemaun AN, Zhu D, Kim JK, Schubert EF, Kim MH, Chung HJ, Yoon S, Sone C, Park Y (2009) Reduction in efficiency droop, forward voltage, ideality factor, and wavelength shift in polarization-matched GaInN/GaInN multi-quantum-well light-emitting diodes. Appl Phys Lett 94:011113. doi:► 10.1063/1.3058687
62. Kim MH, Schubert MF, Dai Q, Kim JK, Schubert EF, Piprek J, Park Y (2007) Origin of efficiency droop in GaN-based light-emitting diodes. Appl Phys Lett 91:183507. doi:► 10.1063/1.2800290
63. Xie JQ, Ni XF, Fan Q, Shimada R, Özgür Ü, Morkoç H (2008) On the efficiency droop in InGaN multiple quantum well blue light emitting diodes and its reduction with p-doped quantum well barriers. Appl Phys Lett 93:121107. doi:► 10.1063/1.2988324
64. Hammersley S, Watson-Parris D, Dawson P, Godfrey MJ, Badcock TJ, Kappers MJ, McAleese C, Oliver RA, Humphreys CJ (2012) The consequences of high injected carrier densities on carrier localization and efficiency droop in InGaN/GaN quantum well structures. J Appl Phys 111:083512. doi:► 10.1063/1.3703062
65. Nakamura S, Krames MR (2013) History of gallium-nitride-based light-emitting diodes for illumination. Proc IEEE 101(10):2211–2220. doi:► 10.1109/JPROC2013.2274929
66. Nakamura S, Mukai T, Senoh M (1994) Candela-class high-brightness InGaN/AlGaN double-heterostructure blue-light-emitting diodes. Appl Phys Lett 64:1687–1689. doi:► 10.1063/1.111832
67. Steigerwald DA, Bhat JC, Collins D, Fletcher RM, Holcomb MO, Ludowise MJ, Martin PS, Rudaz SL (2002) Illumination with solid state lighting technology. IEEE J Sel Top Quantum Electron 8(2):310–320. doi:► 10.1109/2944.999186
68. Krames MR, Shchekin OB, Mueller-Mach R, Mueller GO, Zhou L, Harbers G, Craford MG (2007) Status and future of high-power light-emitting diodes for solid-state lighting. IEEE J Disp Technol 3(2):160–175. doi:► 10.1109/JDT.2007.895339
69. Fujii T, Gao Y, Sharma R, Hu EL, DenBaars SP, Nakamura S (2004) Increase in the extraction efficiency of GaN-based light-emitting diodes via surface roughening. Appl Phys Lett 84:855–857. doi:► 10.1063/1.1645992
70. Narukawa Y, Ichikawa M, Sanga D, Sano M, Mukai T (2010) White light emitting diodes with super-high luminous efficacy. J Phys D Appl Phys 43:354002. doi:► 10.1088/0022-3727/43/35/354002
71. David A, Hurni CA, Aldaz RI, Cich MJ, Ellis B, Huang K, Steranka FM, Krames MR (2014) High light extraction efficiency in bulk-GaN based volumetric violet light-emitting diodes. Appl Phys Lett 105:231111. doi: ► 10.1063/1.4903297
72. Schlotter P, Schmidt R, Schneider J (1997) Luminescence conversion of blue light emitting diodes. Appl Phys A 64(4):417–418. doi:► 10.1007/s003390050498
73. Mueller-Mach R, Mueller GO, Krames MR, Trottier T (2002) High-power phosphor-converted light-emitting diodes based on III-nitrides. IEEE J Sel Top Quantum Electron 8(2):339–345. doi:► 10.1109/2944.999189
74. MacAdam DL (1943) Specification of small chromaticity differences. J Opt Soc Am 33:18–26
75. Murphy TW (2012) Maximum spectral luminous efficacy of white light. J Appl Phys 111:104909. doi:► 10.1063/1.4721897
76. CREE (2014) ► http://www.cree.com/News-and-Events/Cree-News/Press-Releases/2014/March/300LPW-LED-barrier

77. Zhao LX, Thrush EJ, Humphreys CJ, Phillips WA (2008) Degradation of GaN-based quantum well light-emitting diodes. J Appl Phys 103:024501. doi:► 10.1063/1.2829781
78. Amano H, Sawaki N, Akasaki I, Toyoda Y (1986) Metalorganic vapor phase epitaxial growth of a high quality GaN film using an AlN buffer layer. Appl Phys Lett 48:353–355. doi:► 10.1063/1.96549
79. CIE data of 1931 and 1978 (1978) Available at ► http://cvision.ucsd.edu and ► http://www.cvrl.org
80. Juza R, Hahn H (1938) Über die Kristallstrukturen von Cu3N, GaN und InN Metallamide und Metallnitride. Z Anorg Allgem Chem 239:282–287. doi:► 10.1002/zaac.19382390307
81. Lighting Europe (2013) Guide for the application of the commission regulation (EU) No.1194/2012 setting ecodesign requirements for directional lamps, light emitting diode lamps and related equipment. ► http://www.lightingeurope.org/uploads/files/LightingEurope_Guide_-_Regulation_1194_2012_ECODESIGN_Version_1_17_July_2013.pdf
82. MacAam DL (ed) (1993) Colorimetry—fundamentals. SPIE Optical Engineering Press, Bellingham, WA
83. Maruska HP, Tietjen JJ (1969) The preparation and properties of vapor-deposited single-crystal GaN. Appl Phys Lett 15:327–329. doi:► 10.1063/1.1652845
84. Nakamura S (1991) GaN growth using GaN buffer layer. Jpn J Appl Phys 30:L1705–L1707. doi:► 10.1143/JJAP.30.L1705
85. Nakamura S, Senoh M, Iwasa N, Nagahama S (1995) High-brightness InGaN blue, green and yellow light-emitting diodes with quantum well structures. Jpn J Appl Phys 34:L797–L799. doi:► 10.1143/JJAP.34.L797
86. Nakamura S, Senoh M, Mukai T (1993) P-GaN/N-InGaN/N-GaN double-heterostructure blue-light-emitting diodes. Jpn J Appl Phys 32:L8–L11. doi:► 10.1143/JJAP.32.L8
87. Pankove JI, Miller EA, Berkeyheiser JE (1972) GaN blue light-emitting diodes. J Lumin 5:84–86
88. Razeghi M (2002) Short-wavelength solar-blind detectors—status, prospects, and markets. Proc IEEE 90(6):1006–1014. doi:► 10.1109/JPROC.2002.1021565
89. Zhu D, Wallis DJ, Humphreys CJ (2013) Prospects of III-nitride optoelectronics grown on Si. Rep Prog Phys 76:106501. doi:► 10.1088/0034-4885/76/10/106501
90. Soraa (2015) ► https://www.soraa.com/news_releases/32
91. Philips (2014) ► http://www.philips.com/a-w/about/news/archive/standard/news/press/2014/20140509-Philips-and-Green-Sense-Farms-usher-in-new-era-of-indoor-farming.html
92. ► https://en.wikipedia.org/wiki/Phase-out_of_incandescent_light_bulbs

Modern Electron Optics and the Search for More Light: The Legacy of the Muslim Golden Age

Mohamed M. El-Gomati

M.M. El-Gomati (✉)
Department of Electronics, University of York, Heslington, York YO10 5DD, UK
e-mail: Mohamed.elgomati@york.ac.uk

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_6

6.1 Introduction

This chapter is a brief survey of some of the fundamental premises of electron optics with an emphasis on electron microscopy and its relevance to modern life. The reader is introduced to the basic concept of electron microscopy and the parallels to the more familiar optical microscopes that depend upon the use of light optics. Some recent developments in the technique will be surveyed. The UN General Assembly proclaimed 2015 as the international year of light and light-based technologies, where electron microscopy has played and continues to play a pivotal role in the development of efficient, environmentally friendlier alternative light sources to the incandescent light bulb. These developments follow a scientific method of enquiry that had its roots laid down in the eleventh century by the Arab scholar, Al-Hassan Ibn al-Haytham, known in the west by his Latinised name as 'Alhazen'. Ibn al-Haytham also discovered and correctly explained and described a puzzling effect observed in the field of optics; known as spherical aberration. The correction of this lens-related deficiency in electron microscopy is shown to have been fundamental in developing a new class of efficient Light Emitting Diodes (LED) light bulbs. The occurrence of 'spherical aberration' is also equally important in other imaging devices such as telescopes, and an example showing its pivotal role in the Hubble Space Telescope (HST) is subsequently demonstrated.

6.2 Electron Optics

The field of electron optics is concerned with the formation of a fine beam of electrons to use in a variety of applications as listed below. In one such relevant example; electron microscopy, the electron beam is directed to bombard a solid specimen for the purpose of learning more about the properties of such a sample, but with the particular emphasis on obtaining such information at the smallest possible dimensions. This electron-solid interaction results in an array of signals, or by-products, ranging from electrons to photons. The collection of one of these signals has often led to a specialised technique associated with such a signal. This will be explained in more details later on. There is estimated to be about 100,000 instruments around the world that use electrons to map the features of a given specimen.

Another important technical area concerns the use of a focussed electron beam, which is known as electron beam lithography (EBL). In such applications, the smallest device components like diodes and transistors; being small is essential for high-speed electronics, are drawn using small-diameter electron beams for producing the required masks used in transferring a given pattern composed of these components to a semiconductor wafer. Electron beams have also been used in direct writing on devices, in locally functionalising a surface, as well as in electron-beam-induced deposition of materials, to name but a few relevant examples.

In addition to, the use of a fine beam of electrons to map the surfaces of samples, small-diameter energetic electron beams have also been used in welding metals. The instruments employed for such applications share with the imaging instruments a great deal of commonalities, but the emphasis for welding of metallic materials is on producing an energetic spot of electrons with sufficient current to fuse together the parts of the samples being welded.

Another equally important area that also requires the production of a small-diameter electron beam, albeit not as fine as in electron microscopy, is in conventional television sets and in cathode ray tubes (CRT) used in oscilloscopes for scientific applications. Neither of these methods requires the production of a small

beam diameter, because the human eye can only resolve features of the order of ~200 micro metres (μm) (1 μm is one of a millionth of a metre whilst 1 nm is one billionth of a metre).

As can be appreciated from the foregoing discussion, the subject of electron beam optics is vast in nature and almost impossible to cover all its aspects in a non-specialised book chapter such as this. I will therefore confine the material presented to two main parts: the first is an introductory review of some of the more widely used instruments that are employed to spatially map the constituents (elements) of the solids under investigation. The depth of information gathered in this exercise ranges from the top atomic layer of such a sample down to depths of few microns below its surface. Second, I will briefly turn my attention to some of the latest developments in electron microscopy, and specifically the techniques that have been introduced in the last few years to advance this method. This second part will include some important highlights of the use of electron microscopy in relation to light and light-related techniques as a mark of this year's celebration.

6.3 Parallels with Optical Microscopy

In order to better appreciate electron microscopy, it is helpful to draw a parallel with a much older technique that is also used to image a variety of samples; including solids and liquids; the optical microscope. This instrument is perhaps better known and more widely used, given that many of us can claim to have used one either during science-based education or at work. This fascinating instrument has also over many decades seen extensive research and development to optimise it to the state we are familiar with today. Such advances have in many ways paved the way for the development of the more recent technique, the electron microscope, which has somehow been easier and faster to develop, particularly in its early days. There are, indeed, many distinctive similarities between the propagation of photons (light) and electrons, where examples of ray (photon) optics can be used to illustrate and understand electron optics. Parameters such as focal distance, linear magnification and angular magnification, which define the ray optical systems of optical microscopy, can therefore be understood.

In optical microscopy, we observe an important phenomenon: the way photons behave as they cross a boundary between two media of differing refractive indices causes the incident ray to change direction through a different angle from that of incidence. This phenomenon was formulated, and has come to be known to us, as *Snell's Law* (Willebrord Snellius 1580–1626—with René Descartes having also independently reached the same formulation in his 1637 essay, 'Dioptrics.'). However, recent research findings suggest that this relationship was in fact discovered more than six centuries earlier than the reports of Snellius by an Arab scientist, Abu Sa'd al-Alaa Ibn Sahl (940–1000 CE), commonly known as Ibn Sahl, who resided in Baghdad, capital of today's Iraq. The relatively recent discovery of a manuscript by the French Arab historian of science, Rashed [1] has unveiled the work of Ibn Sahl, who developed this relationship while studying the properties of burning mirrors and lenses. Ibn Sahl showed for the first time the refraction of light by glass lenses, as depicted in Fig. 6.1. He went on to describe an elaborate instrument design which could be used to manufacture glass lenses of regular and varying shapes [1]. Astonishingly, this discovery went unnoticed and uncredited for almost 1000 years!

Snell's law describes the relationship between the incidence angle α and the refractive angle β and speeds v_1 and v_2 of a photon as it crosses the boundary of two media of differing refractive indices, n_1 and n_2, respectively. This is shown diagrammatically in Fig. 6.2 and can algebraically be stated as:

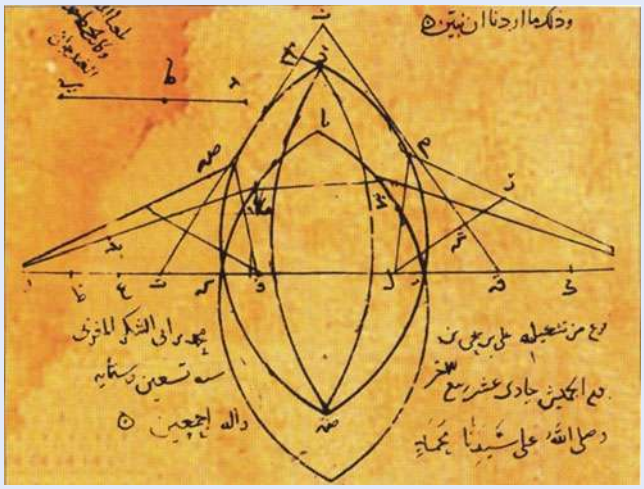


Fig. 6.1 A diagram of the refraction in lenses, by the Arab scientist/mathematician, Abu Sa'd al-Alaa ibn Sahl (d.960CE), who lived in Baghdad, Iraq. The manuscript, dated Thursday 11 Rabi' al-Akhir 690 AH (i.e. 12 April 1291 CE), was discovered and edited by Rashed [1], with permission from the author

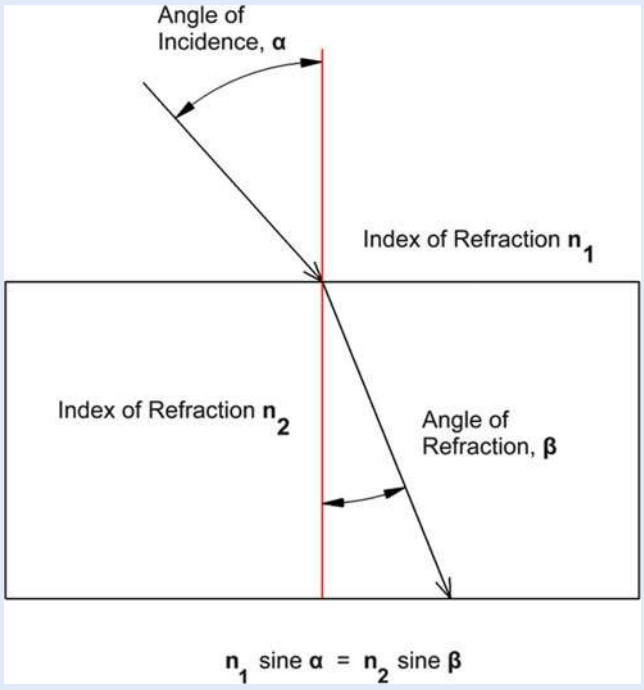


Fig. 6.2 A diagram showing the refraction of light as it crosses a boundary between two media of refractive indices n_1 and n_2

$$\sin \alpha / \sin \beta = n_1/n_2 = v_1/v_2 \tag{6.1}$$

In a simplified way, the behaviour of electrons when moving through a boundary of two areas of differing potentials could also be described by a similar relationship to (6.1) above. Consider the speed ($v_{1,2}$) of an electron of mass (m)

and charge (e) in relation to the potential acting upon it as it moves between two areas of potentials V_1 and V_2:

$$½\, mv_1^2 + eV_1 = ½\, mv_2^2 + eV_2 \tag{6.2}$$

The electric field only applies a perpendicular force on the moving electron. This then causes only the electron perpendicular momentum to be affected; so, as it crosses the (potential) boundary, we can write:

$$mv_1 \sin \alpha_1 = mv_2 \sin \alpha_2 \tag{6.3}$$

The above equation can then be rewritten to get:

$$\sin \alpha_1 / \sin \alpha_2 = \sqrt{(V_2/V_1)} \tag{6.4}$$

Equations (6.1) and (6.4) have a close resemblance with $\sqrt{(V_{1,2})}$ acting as the refractive index. Such close relationships allow one to reflect on the same optical theory of light optics, and to expect to gain similar results up to a certain extent.

This chapter, however, is mainly concerned with the electron microscope family of instruments. Therefore, whilst there has been a great deal of developments and advancements in optical microscopy in the last few decades, they will not be covered here.

6.4 JJ Thomson and His Discovery, the Electron

It would be unfair, if not incomplete, to discuss electron microscopy without mentioning its pivotal component; *the electron*. Many questions arise: so what is this electron, and how was it discovered and by whom? How do we generate electrons for use in microscopy? Are there different electron sources in use today? What are these and is this important?

The discovery of the electron in 1897 by Joseph John Thomson, widely referred to as JJ Thomson, came amidst intensive research activities in a related instrument to the present electron microscope; namely, the CRT. In Cambridge, where JJ Thomson was just appointed as a Professor of Experimental Physics at the Cavendish Laboratory (the Physics Department at the University of Cambridge, UK), he pursued his research into electrical discharges in CRT. The CRT is a vacuum tube which was made of glass in its early days, with a heated filament acting as the cathode opposite a fluorescent screen acting as the anode. The work carried out in one form of the CRT, called the Crookes tube, has shown these rays to cast a shadow on the glowing walls of the glass tubes and on the fluorescent screens opposite and, as such, move in straight lines. Being charged particles, it meant that they could also be deflected by either electrical or magnetic fields (for more details on the history of the cathode rays, see the Flash of the cathode rays—Dahl [2]).

At that time, there were many opinions on the nature of cathode rays as being waves, atoms or molecules. It was JJ Thomson, however, who carefully designed an experiment to measure the electrical charge to mass of the emitted particles, which established them as unique elementary sub-atomic particles. He called these particles 'corpuscles', which later on were given the name 'electrons' by Fitzgerald as a result of combining the words *electr*ic and i*on* [3]. In addition, such particles, which can be accelerated, have a wavelength which is shorter than that of light photons by up to 100,000 times (depending on the electron's speed). These properties led researchers later on to accelerate a focussed beam of electrons to illuminate the surface of a specimen—as light is used in optical microscopy—and

hence develop electron microscopy as an imaging tool. As a result of this, a resolution imaging limit of about 50 pico-metres (pm) was estimated for electron microscopy, which is yet to be realised, in comparison to an upper limit of some 200 nm for traditional light microscopy. These exciting prospects have encouraged research into developing imaging tools using energetic electron beams. It was in 1931 that Ernst Ruska and Max Knoll [4] finally succeeded in demonstrating the first working electron microscope.

There are two major types of electron microscope; the first of which uses a high energy beam of electrons to penetrate thin samples, and the transmitted electrons are used to form an image of that sample in what has become known as the transmission electron microscope (TEM). The electron beam energy in this case ranges from 50,000 V up to 1,000,000 V, but most widely used electron energies in TEMs today are in the range 100,000–300,000 V. In the second type, known as the 'scanning' type, the incident electron beam is normally in the range 100–30,000 V. These electrons are arranged to impinge the solid surfaces, and the reflected electron signal is used primarily to map the surface topography of the said sample. This type is known as the scanning electron microscope (SEM) and is the most widely used type in industry and academia alike. The major difference between the two instruments is in the collection of the signal from each, in the TEM the signal is of the transmitted electrons through the thin sample, whilst in the SEM the signal is normally of the reflected electrons (i.e. scattered back off the sample's surface). Both types use electrostatic and electromagnetic *electron lenses* to control the electron beam energy, but more importantly for the SEM is to focus it in the smallest possible spot which, when scanned, could be used to form an image of the area the electron impinges.

These electron optical *lenses* are analogous in function to the glass lenses of the optical light microscope. The remainder of this chapter will review the principle of the electron microscope, its major components, and provide a short overview of its applications, particularly in relation to light-related technology, before finally concluding with future trends in electron microscopy.

6.5 The Principle of Electron-Solid Interaction

Imagine an electron beam of an infinitesimally small diameter is incident on a solid surface, as depicted in Fig. 6.3 below. As these electrons penetrate the solid, they will interact with the sample's constituent atoms. This interaction is referred to as 'electron scattering' and is further divided into two categories: (a) an inelastic scattering, where the incident electron gives up part of its energy as a result of its collision with the solid's atoms; and (b) an elastic scattering, which causes the electron to change its direction of travel with almost no energy loss. It is the first scattering type, however, that gives rise to the signals enumerated in Fig. 6.3, whilst the second type is normally what determines the shape of the interaction volume of the incident electrons within the solid under study, as depicted in Fig. 6.4. The shape and size of the interaction volume depends on the incident electron energy, its angle of incidence with respect to the surface and on the average atomic number of the sample under study. If one concentrates on the emitted electron signal alone and plots their number against their energy, one would in principle collect a distribution similar to that shown in Fig. 6.5.

The collection of one of the signals resulting from the interaction depicted in Fig. 6.3 has over the years resulted in a specific class of instruments reflecting the type of information gathered from such interaction. For example, if one collects the resulting X-ray photons of a given element making up the solid, the collected image would be a map of the distribution of such an element in the solid, normally

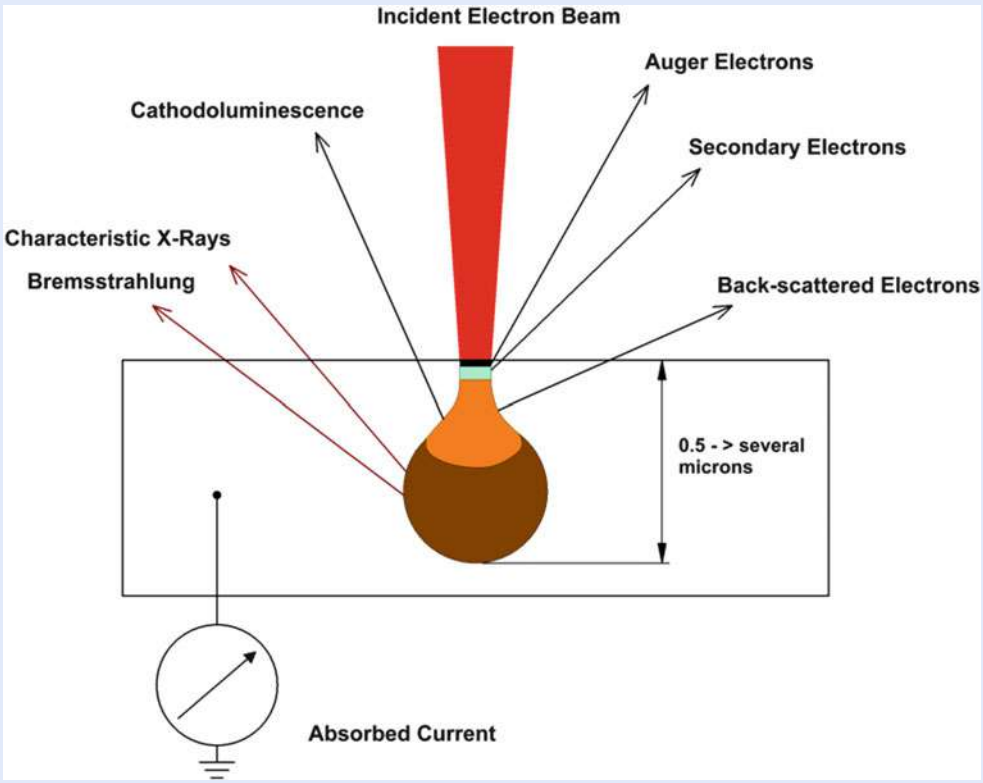


Fig. 6.3 A schematic depicting electron-solid interaction on a solid sample, and the various signals that result from such interaction

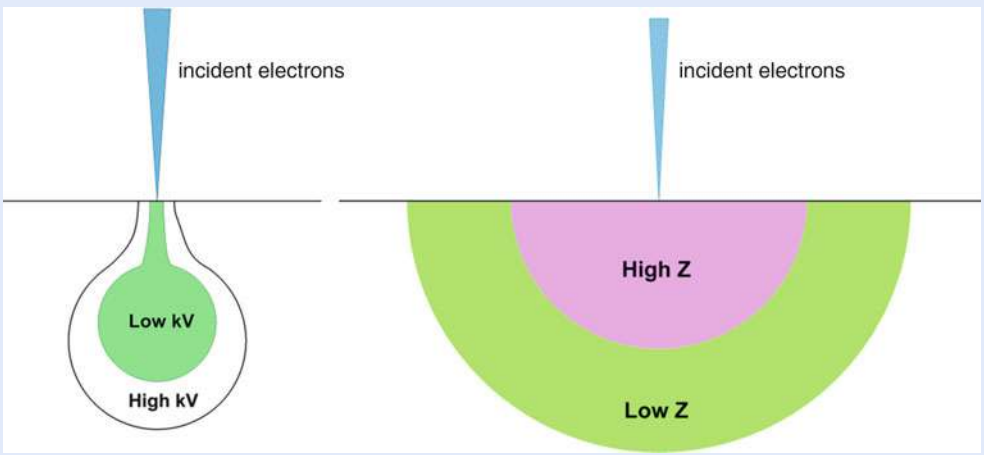


Fig. 6.4 A schematic showing electron-solid interaction where in (**a**) it shows the difference in the overall volume as a function of the incident beam voltage, whilst in (**b**) it shows the difference in volume between high and low atomic number materials

to a depth ranging from ~0.05 μm to few μm depending on the incident electron energy and the sample's atomic number. This technique is referred to as the electron microprobe analysis (EPMA) and is heavily used in material science research and applications. If, however, one uses the Auger electrons of a given element, then these would normally produce a map of the distribution of this element on the very top atomic layers of such a sample—a technique known as

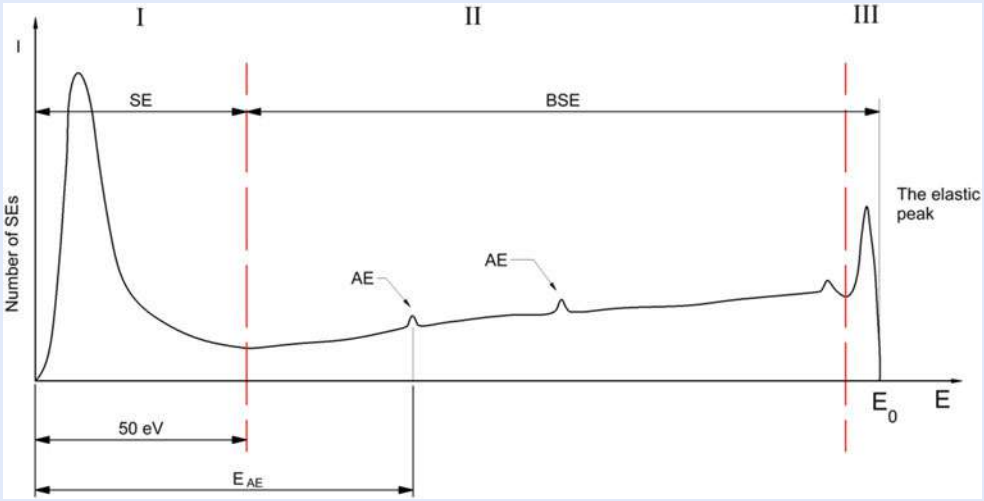


Fig. 6.5 Electron spectrum of a typical distribution in energy and number of electrons that exit the surface of a solid which is bombarded by a beam of energetic electrons. *AE* Auger electrons, *SE* secondary electrons, *BSE* backscattered electrons

Scanning Auger Electron Microscopy (SAM). Auger electrons are an alternative to X-ray emission from atoms and also, like the X-ray photons used in EPMA, are material specific, i.e. almost the equivalent of finger prints of the elements [5]. Whilst X-rays originating a few microns below the surface could still be collected, the collected Auger electrons come from only the top few atomic layers of the solid under study (i.e. the depth of information of the X-rays could be >100 times or more greater than that of the Auger electrons). The collection of other signals would give rise to techniques associated with such other signals.

However, the main and most popular type of electron microscope is the SEM, where use of the reflected low-energy 'secondary' electrons, which mainly have an energy of <50 eV (refer to Fig. 6.5), is made to map the topography of the solid surface. One important application of the SEM is in the semiconductor industry, both during the development of the integrated circuits and the various electronic devices constituting it but, equally as important, during their production (i.e. in what has become known as fabrication lines). In the latter case, there are normally two types of microscope in use. The first is referred to as a *Critical Dimension Scanning Electron Microscope* (CD-SEM). This type of instrument is normally used on semiconductor fabrication lines for quality control, with the main function of measuring the dimensions of some components on the circuits/devices being produced. It is also interesting to note that such inspection is usually made automatically and lasts no more than few seconds per test, making the inspection of some 10–20 areas on a full 300 mm diameter wafer last no more than 1 min.

The second type is normally also used for quality control, with the main purpose being searching for 'defects' or foreign particles appearing on the wafer which could cause the ultimate failure of the circuit built on or next to it, for example, where a particle is found lying between two active parts of the circuit and it unintentionally connects them. These defects can be due to anything ranging from small particles that find their way into the production line as a result of poor practice to failure in procedure. Sometimes these also appear as a result of the use of poor quality or incompatible materials in the fabrication process including, for example, the 'pure or de-ionised' water used for washing the wafers. This class of instruments is referred to as '*defect review SEMs*', which are normally equipped with a number of elemental/chemical detectors used to analyse such defects. The SEM is also equally vital in biological and other physical sciences, such as physics

and engineering. More discussion will be devoted to the SEM later on. In recent years the development of compact and sometimes novel X-ray and various detectors has enabled these methods to be used as add-on techniques to the electron microscope, thus enhancing the instrument's analytical capability and widening its use.

If, on the other hand, the sample under study could be made in the shape of a thin enough section to allow a beam of energetic electrons to penetrate it and be collected from the other side, then a whole new array of signals would result in allowing one to gain more information from such a sample on atomic dimensions. This class of instruments is referred to as TEM. The working of the TEM, however, is quite subtle and different than that of the SEM. To understand the basic principle of TEM, let us investigate the fate of some energetically incident electrons on a thin sample. The transmitted electrons will pass through the thin specimen but, in so doing, will be subjected to one of three possible mechanisms:

1. To pass through with no scattering,
2. To pass through with some angular deflection (elastic scattering) and
3. To lose some energy as it passes through (i.e. inelastic scattering).

The above three possibilities are normally a function of the energy of the incident electrons and the arrangement of the sample's atoms. How we see the varying contrast in the obtained TEM images is quite interesting though.

To better appreciate the underlying principle behind the operation of the TEM, it is useful to consider the following simple example. Imagine a hypothetical specimen made out of four regions of carbon and lead as depicted in Fig. 6.6 below (where lead is a much higher atomic number material than carbon). If 100 energetic electrons are incident perpendicularly on the surface of this sample and the transmitted electrons are classified according to them being scattered (i.e. change direction of travel) through a small angle of 0.5° or more in comparison with those which pass through without suffering any scattering, or less than 0.5°, then the number of transmitted electrons varies in terms of their scattering

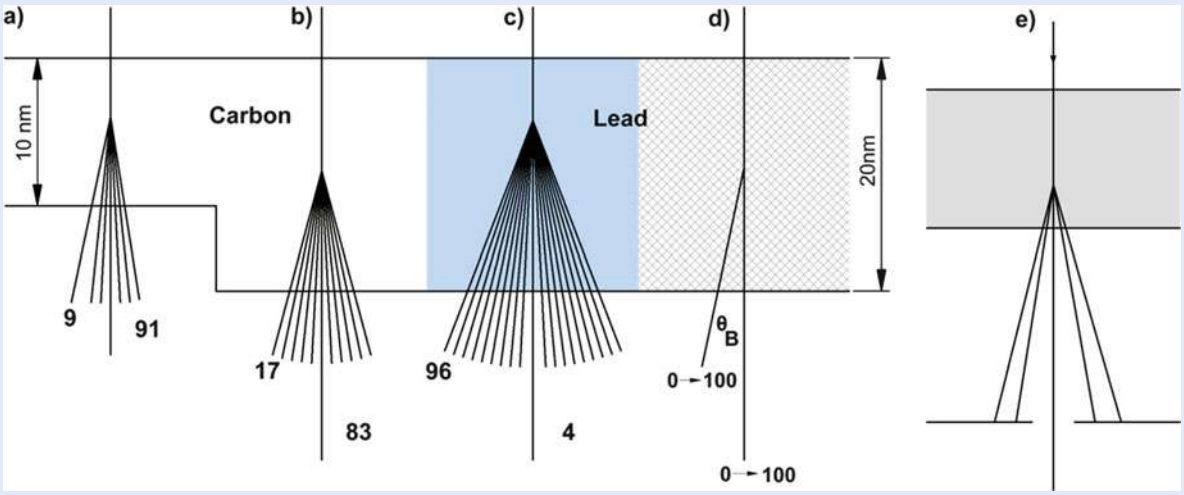


Fig. 6.6 A hypothetical sample used to illustrate the principle of transmission electron microscopy by following the fate of 100 energetically incident electrons, and counting those that scatter through an angle of more than 0.5°. (**a**) carbon film 10 nm thick of randomly distributed atoms, (**b**) same carbon film but of 20 nm thickness, (**c**) 20 nm film of randomly distributed lead atoms, whilst (**d**) is 20 nm thick lead but of regularly distributed atoms (i.e. crystalline), (**e**) the objective aperture which sits below the sample in the TEM to stop electrons scattering through more than 0.5° (adopted from [7]) from passing through and hence contributing to the used signal

angle in a very interesting way. In part (a) of the sample, nine electrons will scatter through 0.5° or more, in part (b) about 17 will scatter, while part (c) shows a much larger number of about 95 electrons scattering by more than 0.5°. However, in part (d) the number scattered depends on the angle that the incident beam of electrons makes with the sample atoms in what is known as 'Bragg diffraction' [6]. If a small aperture (i.e. a hole cut in a metal plate) is placed below the specimen such that any electrons scattered by 0.5° or more are stopped by this plate whilst those scattered by less than this angle go through, then a number of different 'in-value' electron signals could be collected below the small aperture. The position of this aperture is therefore the key part in the working of the TEM. The detectors used in the TEM also vary from principally detecting the crystallinity of the sample, via the distribution of its atoms; or they determine its atomic number, via the measurement of very small energy losses that the incident electrons suffer in passing through the thin sample.

The TEM is a more complex instrument to manufacture and operate than the SEM. Moreover, its price is normally several times more than that of the SEM. The number of TEM instruments worldwide is perhaps less than 10 % of the installed user base of SEMs. It should, however, be understood that the information gathered from either instrument normally complements the other rather than being an alternative to it.

6.6 The Basic Components of Electron Microscopes

Figure 6.7 depicts a schematic of the two most widely used types of electron microscope: the SEM and the TEM. The major components of either instrument could be divided into the following parts: the electron source, the probe-forming column, the specimen chamber and finally the detectors. A brief coverage of these components will be given below; however, for more detailed discussion of the modelling of these components the reader is referred to [8]. It should also be noted that most, if not all, modern instruments are computer controlled, but this will not be covered here.

6.6.1 The Electron Source

The first type of electron source used in electron microscopes consisted of a heated filament made out of a thin tungsten wire, a similar material to that used in conventional light bulbs. Over the years more and more electron source types have been developed to address the fundamental problem of using the highest possible number of electrons per unit area per unit angle of emission; a concept referred to as the 'source brightness'. The higher the brightness value, the smaller the focused point that can be formed with the same number of incident electrons. Table 6.1 below summarises the currently available and widely used types of electron sources and their relative properties. The development of field electron emitters as high-brightness electron sources in the last 40 years or so has in particular moved both types of electron microscope discussed here and other probe-forming systems closer towards realising their ultimate resolving power.

The basic principle of electron emission from the various sources is depicted in Fig. 6.8. The important property that underpins all of these is the material's work function, a property that dictates how much energy is required to liberate an electron from its bound state in the material's atom. For thermionic sources, electrons are liberated by heating the source to 'boiling point', the consequences

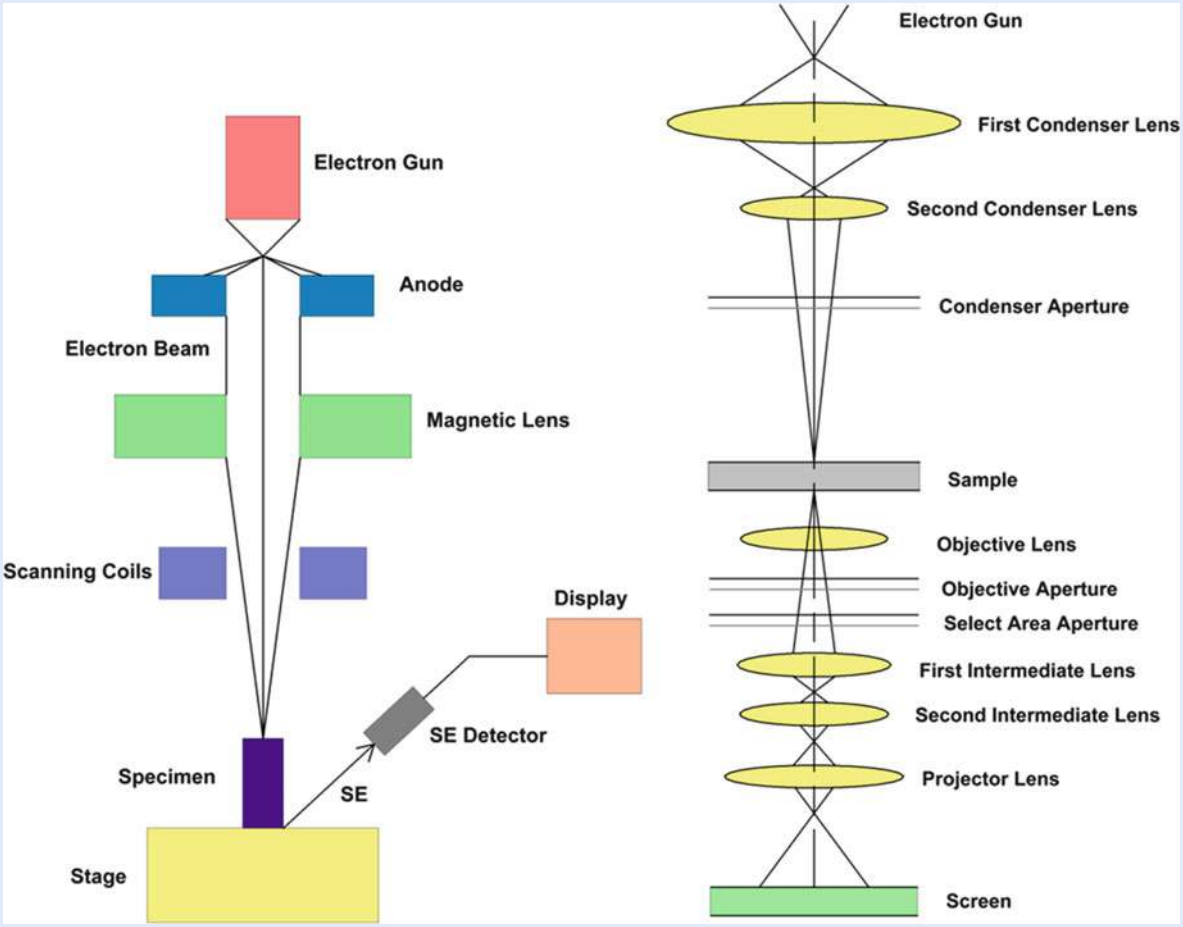


Fig. 6.7 A schematic of the major components of the electron microscopes (**a**) SEM and (**b**) TEM. Note that most microprobes and EBL instruments share the same configuration as the SEM

Table 6.1 A comparison of the characteristics relevant to electron optics for some of the most widely used electron sources (York Probe Sources Ltd, ► http://www.yps-ltd.com/)

Emitter type	Thermionic	Thermionic	Schottky FE	Cold FE
Cathode material	W	LaB_6	ZrO/W (100)	W (310)
Operating temperature (K)	2800	1900	1800	300
Effective source radius (nm)	15,000	5000	15 (*)	2.5 (*)
Normalised brightness (A/cm^2 sr kV)	1×10^4	1×10^5	1×10^7	2×10^7
Energy spread @ cathode (eV)	>0.59	>0.50	0.5–0.8	>0.23
Beam noise (%)	1	1	1	5–10
Operating vacuum (mbar)	$<1 \times 10^{-5}$	$<1 \times 10^{-6}$	$<1 \times 10^{-9}$	$<1 \times 10^{-10}$
Typical cathode life (h)	~100	~1000	>10,000	>10,000

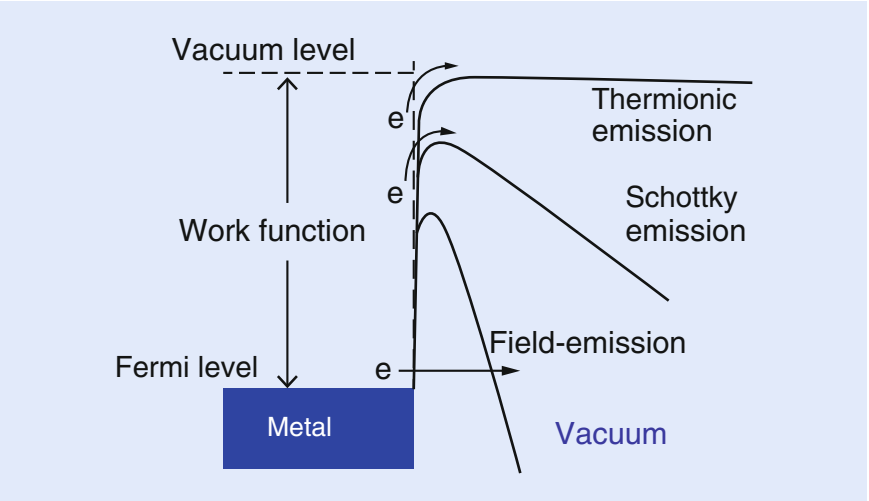


Fig. 6.8 A simplified schematic of the electron potential energy at the solid–vacuum interface showing the modification (*bending*) of the energy barrier as a result of the application of the electric field. For a more detailed description of this phenomenon as applied to various sources, see [10]

of which make the lifetime of this type of source quite limited to mostly no more than few 100 h of use.

The work function of a material is measured in units of electron volts (eV). An eV is the amount of energy gained or lost by the charge of a single electron that is moved across an electric potential of 1 V. For tungsten (W), the work function is about 4.6 eV. The work function acts as a barrier between the electrons in the metal and the vacuum outside the metal. There are a number of materials and combination of materials that have a lower work function than W. One of these is a compound known as lanthanum hexaboride (LaB_6), which has a barrier height of only 2.8 eV; and, when heated to a moderate temperature of about 1800 K, the electrons have enough energy to jump over the barrier thus enabling them to be used as an electron source. The lifetime of this type of source is of a factor of $\times 3$–5 longer than the heated W filament; and their brightness is also more than $\times 10$ higher. The vacuum environment for the operation of this source is, however, more stringent than that of the W counterpart, normally in the region of 10^{-7} mbar.

Field electron emission [9], on the other hand, relies on a totally different mechanism, which exploits the unique wave-particle duality of the electron (see Chap. ▶ 1 for more explanation). This wave nature of the electrons, when utilised, has allowed scientists to develop electron sources with a brightness value of more than 100,000 greater than that obtained from the early tungsten filament counterpart (see Table 6.1 above). In this method, the electron source is made of a very sharp needle, often less than 1 μm in diameter. When this is subjected to a high electric field, $>10^9$ V/m, the barrier between the electron position in the metal and that of the vacuum level becomes so thin (refer to Fig. 6.8 [10]) that electrons can tunnel through it and are liberated into the vacuum. Depending on the electron emitter's diameter and its position relative to an electrode positioned in front of it, called the extractor, the high electric field needed for the electrons to tunnel through can be achieved by applying a moderate voltage value between the emitter and the extractor of less than 5000 V.

Figure 6.9 depicts the various materials and configurations of electron sources currently in use in electron optical instruments which give brightness values spanning a wide range, but which for brevity we will not cover here. Suffice it to say, however, that over the last few decades, field electron emitters, and

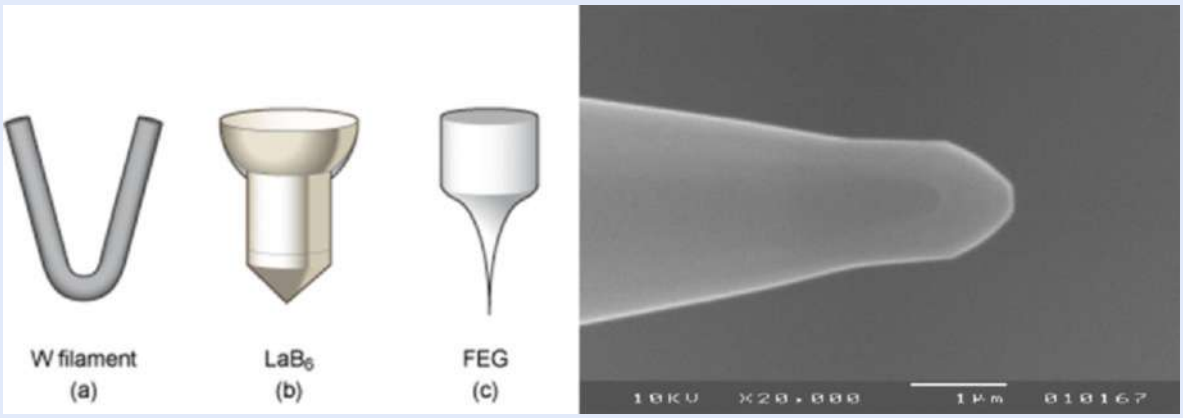


Fig. 6.9 The various electron sources in use: (**a**) the thermionic tungsten (W) filament; (**b**) the low barrier material of lanthanum hexaboride (LaB_6); and (**c**) the field electron emitter (see, for example, MyScope at ▶ http://li155-94.members.linode.com/myscope/sem/practice/principles/gun.php). (**d**) Is a high magnification electron microscope image of a typical Schottky field electron emitter (York Probe Sources Ltd, ▶ http://www.yps-ltd.com/)

particularly those which are known as thermally assisted or 'Schottky' field electron emitters, have become so popular and relatively easy to use that at least half of the modern SEMs and TEMs are manufactured using this class of electron sources. But the heated tungsten filament, in spite of its clear disadvantages, is still expected to be with us for many years to come because of its relatively low cost and ability to function in poor vacuum conditions.

Recent advances in electron source technology are currently directed at developing small-diameter sources of novel materials in pursuit of even higher brightness. Such materials include carbon nano-tubes (CNT) and nano-rods of conducting materials in general. The CNT is a sheet of carbon atoms rolled to form a cylinder of one or several layers. The diameter of this cylinder ranges from few nm to several tens of nm (see Fig. 6.10 for illustration). However, in spite of the potential advantages of such sources in terms of their superior brightness and small source size ([14], it is expected to be sometime before we will see such a source in use due to the great technical challenges in reliably producing and using a stable electron emitter of this type.

6.6.2 The Probe-Forming Column (Electron Lenses)

The probe-forming system consists of a number of 'lenses' that act to form an image of the source of electrons which is used in a focussed spot of electrons. The basic underlying principle of an electron lens is a structure that controls the electron trajectory, ultimately forcing them to converge to a focussed spot. Two types of lens are used in probe-forming systems: an electrostatic one where an electric field is formed by the application of suitable voltage to a specially shaped metal structure which the electrons travel through; or a magnetic lens where an electric current passes through a coil enclosing a carefully shaped iron structure, thus producing a magnetic field that acts on the electrons in a similar fashion to the electrostatic lens. The electrons that pass through such lens structures could then change their path in a diverging or converging manner to a point somewhere away from its starting position (i.e. to form a focused image of the emitted electrons away from its starting point). Such a focused spot could then be used to scan the

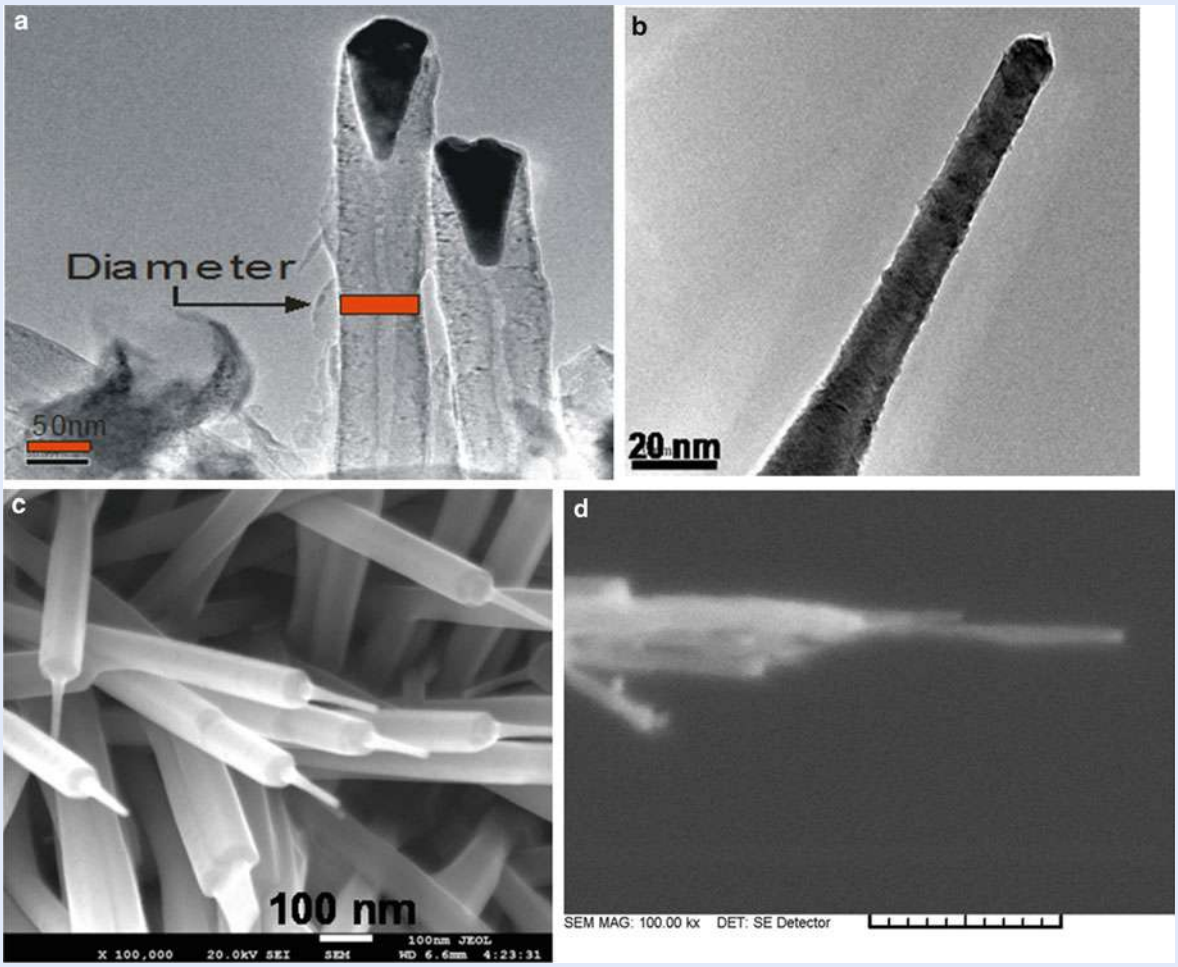


Fig. 6.10 (**a**) TEM image of the end-form of a CNT showing the nickel seed used to grow the CNT. (**b**) TEM image of the ultra-fine tip (nano-needle) of the as-grown In-doped ZnO nano-pencil with a diameter in the range of 13–15 nm. (**c**) SEM image showing the morphology of In-doped ZnO nano-pencils on silicon substrate prepared by thermal evaporation process. The image shows a useful nanostructure made of two parts: ultra-fine tip connected to a base of ZnO nano-rods which could be useful as an electron source. (**d**) Shows a CNT grown on top of a sharp tungsten emitter and the bar measure 1 μm. All of these sources are from the author's research group at the University of York [11–13]

surface of a solid in a raster fashion (i.e. moving the spot across the surface in a straight line and then moving quickly back to the start point and scanning across the next line down) where, at each pixel of the raster, an interaction similar to that depicted in Figs 6.3 and 6.4 takes place. Note that the focussed spot could be smaller or larger than the area where electrons are emitted from depending on the source of electrons used and the applications for which the probe is intended. Note also that the choice to use either an electrostatic or a magnetic lens, or indeed a combination of both, is largely dictated by the intended applications. However, it should be noted that magnetic lenses are more favourable for high resolution applications because of their favourably smaller aberration coefficients (see below for detailed discussion on lens aberrations). The focussed spot diameter (d_T) is therefore composed of a number of contributing sources where their diameters can be simply expressed as the original source of emitted electrons (d_o), the chromatic

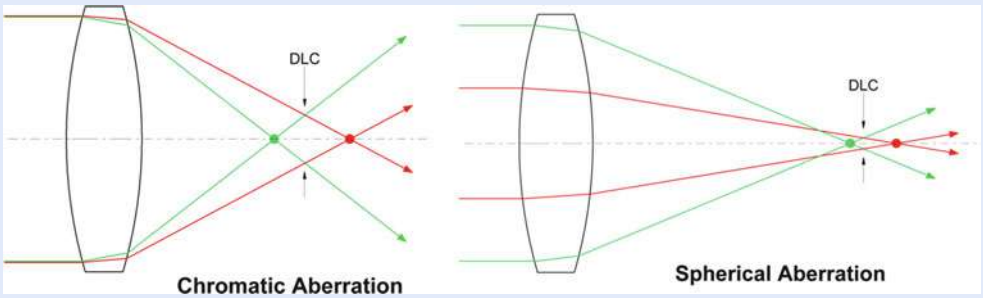


Fig. 6.11 (**a**, **b**) Schematic diagram illustrating chromatic and spherical aberrations and the disk of least confusion (DLC)

aberration (d_C), spherical aberration (d_S) and diffraction effects (d_D). These contributions could be added in the following manner:

$$d_T = \left(d_o^2 + d_C^2 + d_C^2 + d_D^2\right)^{0.5} \tag{6.5}$$

The lenses within an electron microscope behave in a similar fashion to the glass lenses of optical microscopes and these, too, suffer from a number of fundamental properties (or deficiencies) that adversely affect the quality of the focussed image. These are referred to as lens aberrations which basically blur the focussed image, and would therefore require a correction procedure to resolve. Some of these are inherent to electron microscopy, such as the chromatic aberration, due to the nature of the emitted electrons and the type of source used having a range of different initial energies as they leave the surface of the electron source. The effect of such varying energies is that the focussed electrons would correspondingly land, for each electron energy, at a slightly different distance away from the point of emission, as depicted in Fig. 6.11a. Table 6.1 above lists the energy spread typical of the various electron sources in use. Modern electron microscopes, particularly in the case of TEMs, have started to use an energy mono-chromator to correct for the chromatic aberration, but these are not yet commonly used due to the cost they add to the microscope. It is clear from Table 6.1 that field electron emitters have a smaller energy spread than thermionic sources and these give this class of electron sources a clear advantage for use in electron microscopy and particularly so in the case of the SEM, which only use chromatic aberration correctors in the top of the range instruments.

Another optical effect inherent in all cylindrical or spherical lenses (and mirrors) is the spherical aberration, which in the present case is caused by electrons focussing at different points, depending on how far the electrons are incident from the centre of the lens as it passes through it. The reason for this is that for each of the emitted electrons going through the lens's hole when these are projected back would subtend a range of angles at the point of emission (the source). Again the positions of the focussed electrons would converge differently for each electron trajectory, as the electrons travel through regions of different potentials, and hence will obey the formula given above (Eq. (6.4)). As a result they will focus at a range of distances away from the point of emission. The extent of this focal distance away from the point of emission (or the ideal focal point of such a lens) is a function of the angles subtended, i.e. the size of the lens' hole. This in turn means that the larger the aperture-hole diameter, the worse is the effect. Spherical aberration (C_s) is a feature of all round lenses which causes image distortion and limits the ultimate microscope resolving power. In practice, however, is that the various foci of the electrons past the lens will go through a minimum called the 'disk of least confusion' (DLC), which is normally assumed

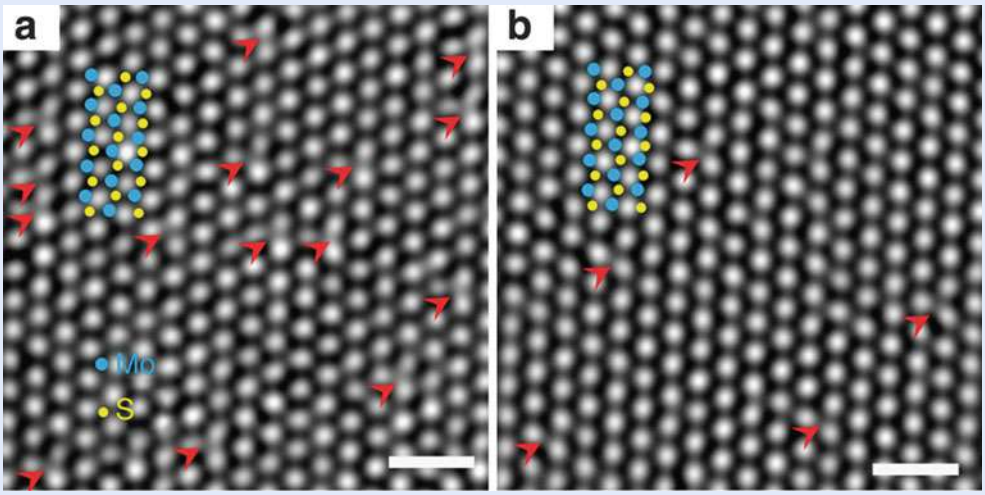


Fig. 6.12 The effect of correcting spherical aberration in a TEM [16]. Note that the *red arrows* refer to points where the separation between the imaged atoms becomes clearer after the spherical aberration correction

as the focal point of the lens. The spherical aberration is schematically illustrated in Fig. 6.11b.

It is interesting to note that spherical aberration was first noted by the Arab scientist, Ibn al-Haytham (d1040) in his magisterial work on optics, 'Kitab al-Manazer' (see Chap. ▶ 2). Whilst this defect was known early on during the electron microscopy development, an effective resolution was only realised and proposed in 1997 (see [15] for a full account). Again, the effect of spherical aberration tends to be more obvious in the TEM case which, when corrected, allows one to resolve structures a small fraction of a nanometre. Its correction in the SEM is also equally important but only applicable in the top of the range instruments. This has in great part been due to the use of field electron emitters, which has made it easier to reduce its effect in SEMs. Figure 6.12 shows the effect of correcting spherical aberration in a modern TEM.

Spherical aberration is an effect typically found in imaging instruments, such as cameras and telescopes. Its presence compromises the quality of the resulting image with potential catastrophe, as the scientists in charge of the Hubble telescope found [17]. The HST project cost several billion US dollars and took several years to complete. The first images collected were of very poor quality, as shown in Fig. 6.13a. The fault causing such a poor image was identified to be due to spherical aberrations of one of the telescope's mirrors and a repair mission was launched to fix it. It consisted of a number of astronauts, one of whom was an experienced experimental scientist who was able to walk in space to repair the fault. Figure 6.13b shows the effect of correcting the spherical aberration on the quality of the images obtained.

The SEM, on the other hand, has also seen a great deal of advancement to improve the quality of the obtained images as well as increasing its resolving power. This has largely been not only due to the use of Schottky and cold-field electron emitters as electron sources, both having lower energy spread and much higher brightness than its conventional thermionic counterpart, but also due to a range of efficient electron detectors that have recently been developed. As a result most modern SEMs offer high resolution imaging in the range of 1–2 nm. These field emission sources have particularly benefited the use of low voltage imaging where most modern instruments operate down to few 100 eVs and sometimes even less to obtain an image resolution of the order of few nm. This high resolution

Fig. 6.13 Hubble Space Telescope images before and after the STS-61 mission [17]

low-electron energy mode is particularly important in imaging biological and novel and radiation-sensitive materials, as encountered in the field of nanotechnology.

By reducing the incident electron beam energy, the interaction volume between the incident electrons and the specimen also reduces and becomes much closer to the specimen surface, as seen in Fig. 6.4 above. This extends the use of the SEM to previously unchartered areas. One recently developed field of application is in imaging doped regions of semiconductors, which have traditionally been almost impossible to image using conventional high-electron energy imaging (>5 keV). This is because the current secondary electron detectors used in the SEM are incapable of resolving differences in the atomic number (Z) of the constituting elements of a sample of less than $Z = 1$. However, the amount of doping material used in semiconductor devices is normally less than 0.001 %; and yet by imaging with electron energy of less than, say, 2000 eV one could differentiate between regions of a sample like silicon, which are differently doped, as depicted in Fig. 6.14. Doping of semiconductors is the crucial step in making electron devices, such as the p-n junction diode; the basic building block of integrated circuits. This low-energy imaging mode is proving valuable in the fabrication lines in the semiconductor industry, where its use is in quality control. The explanation of the mechanism that gives rise to such a contrast is outside the scope of this chapter, and the reader is referred to recent articles discussing this phenomenon [18, 19].

A recent development in SEM technology to cope with the limited space available in modern laboratories is the advent of the table-top electron microscope. A number of microscope manufacturers now specialise in producing such instruments. It is estimated that many thousands of such microscopes have been produced and sold to date. It is expected that further developments will continue to be exerted in this endeavour. One such development from the author's laboratory, depicting a small-size electron column, is shown in Fig. 6.15. This column also contains a novel electron detector inside the column [20]. One other parallel development has been in using silicon wafer technology to make the various microscope column components, which include the lenses, the deflectors and the stigmator correctors, as depicted in Fig. 6.16 [21]. The whole column in the latter case measures only a few mm in height making the whole microscope with its control electronics, vacuum system and specimen chamber comparable in size to a laser printer. This SEM is now commercially available.

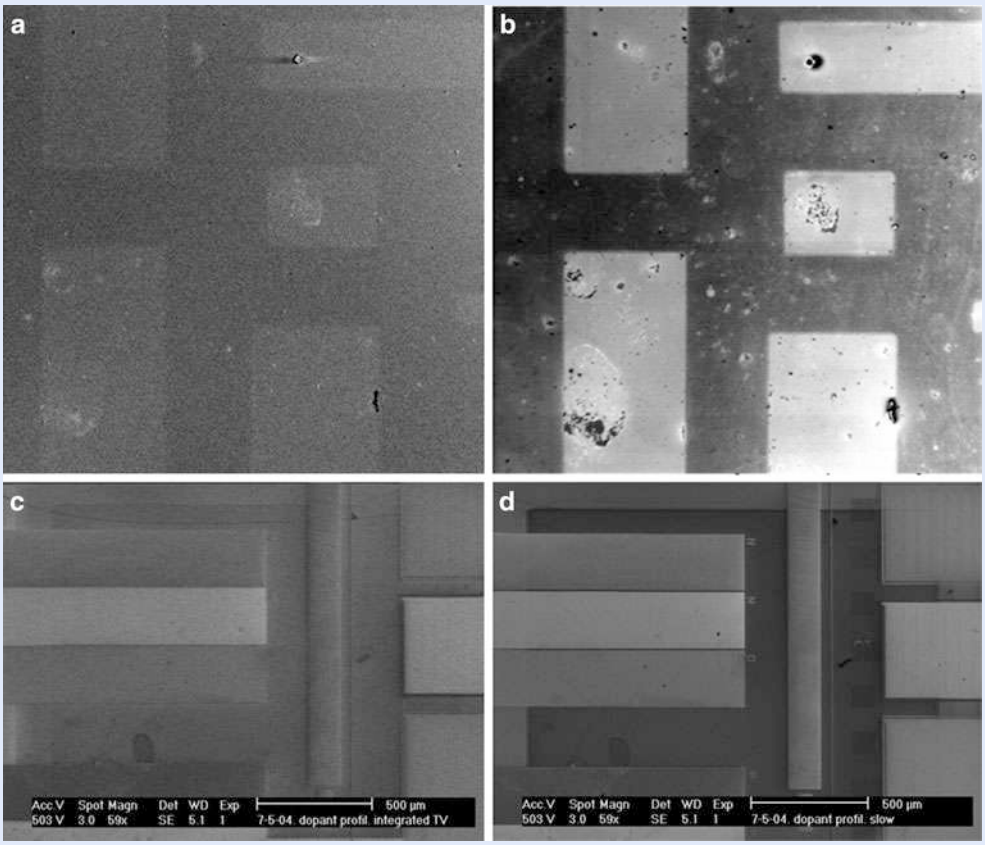


Fig. 6.14 The imaging of doped semiconductors with low-voltage electrons (Image (**a**) collected at 6 keV) and (**b**) collected at only 2 eV). The contrast of the p-regions, which are the bright areas, is increased as the beam energy is decreased. Note in particular the appearance of some small surface particles (contaminants) when imaging at such low-electron energies. In images (**c**, **d**) another effect is shown which is related to the scanning speeds the images are collected (**c**) at high speeds (TV rates) and (**d**) at slow scanning speeds of only 2–5 s per frame [18–20]

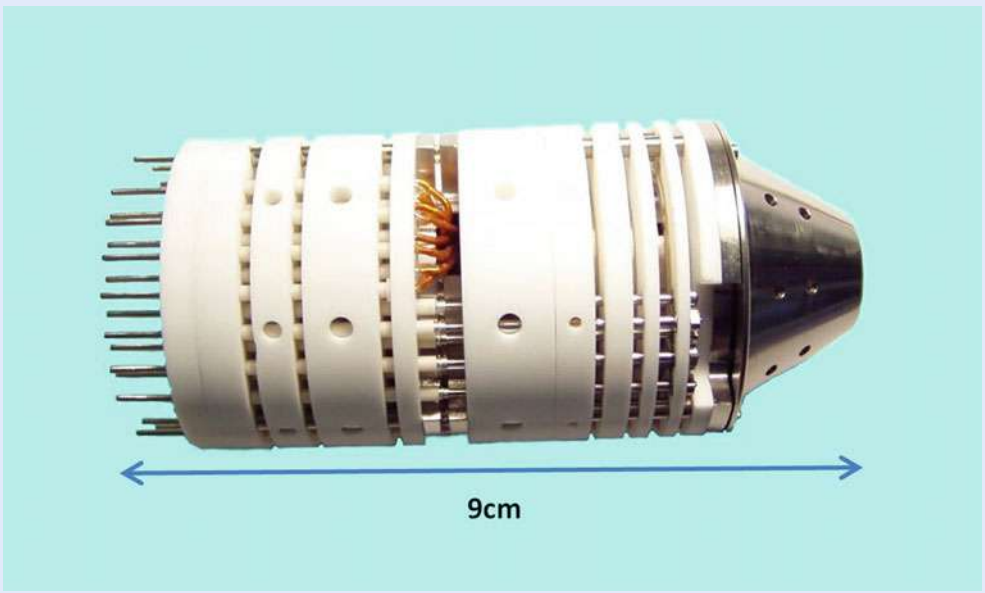


Fig. 6.15 A whole electron column from the author's laboratory, developed for imaging at ultra-low voltage use, down to 1 eV. It contains all the necessary components to form an image, a Schottky field emission source, the electron lenses and deflectors as well as a novel in-lens electron detector [20]

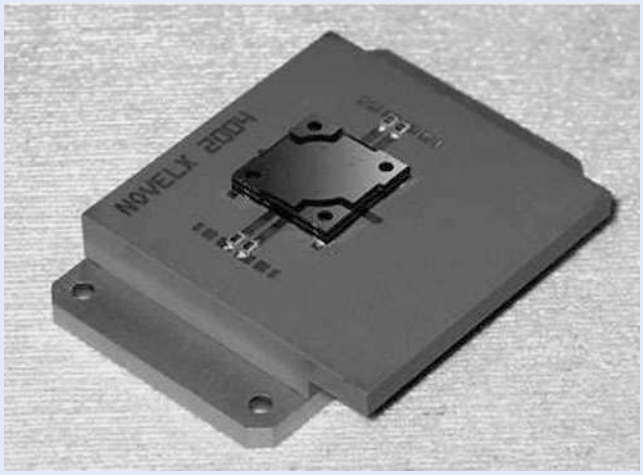


Fig. 6.16 A fully assembled miniature Si-wafer-based column. The dimensions of the whole column are 30 × 30 × 9 mm [21]

The Specimen Chamber

The specimen chamber of the electron microscope is an airtight metal vessel that is pumped to a pressure of the order of 10^{-5} to 10^{-6} mbar (standard sea level pressure is defined as 100 mbar). Again, the size of the specimen chamber differs greatly between the TEM and the SEM. The former can only accommodate samples measuring a few mms in diameter, where the specimen to be studied occupies only a small fraction of this size. The SEM, on the other hand, could accommodate samples a few cm in diameter or even larger, as in the semiconductor industry where full wafers measuring up to 30 cm in diameter are normally used in critical dimension type SEMs. Often, however, in material science, biology and general engineering, samples seldom measure in excess of several mm in diameter. Therefore most specimen chambers in SEMs measure between 20–40 cm to accommodate the various samples from these disciplines.

6.6.3 The Detectors

It is also customary that the specimen chamber, particularly for the SEM, accommodates a number of detectors to capture the various signals emitted from the sample being studied, as depicted in Fig. 6.3. Associated with the choice of a given detector is a requirement to control the total pressure around the sample. For example, if one needs to map the distribution of a given element on the top few atomic layers of a solid sample (as in Auger electron spectroscopy [4]), then the pressure around the sample should be much lower than if the signal being captured is only to reflect the surface topography (i.e. using the low-energy electrons depicted in Fig. 6.5). In the former, a pressure in the region of 10^{-9} mbar or lower is needed, whilst for the latter the environment around the sample could be as high as 10^{-5} mbar. There is now a class of instruments, referred to as environmental SEM (ESEM), where the pressure is even higher than conventional SEMs ($>10^{-2}$ mbar). This is to allow for the inspection of insulated or biological specimen, including glass, wood and even meat.

The development and use of energy-dispersive X-ray detectors (referred to as either EDS or EDX) has, however, given electron microscopes, and particularly the SEM, an added dimension in analytical power. Traditionally, X-ray analysis of solid samples has been carried out using dedicated instruments that employ wave-dispersive type detectors (WDS) (i.e. where the emitted X-rays are detected according to their wavelength for the WDS rather than according to their energy, as in EDS). Detecting the emitted X-rays was the task of the first electron microprobe for use as an analytical tool, and it is referred to as the electron probe microanalyser (EPMA) [22]. The EPMA is a much more complicated instrument than the SEM due to the special X-ray detectors used, but of course it has more powerful analytical abilities, too. In addition it has a higher X-ray resolving power than the EDS type and thus enables smaller amounts of materials to be detected. However, EDS is cheaper and simpler to operate. It is estimated that up to about half of all electron microscopes in the world are equipped with an EDS detector, which is mainly used for qualitative elemental analysis with detection limits of around one atomic %.

There are many other types of detectors that are also used to gain more information from the sample under investigation, and the semiconductor industry leads in this respect. Some of these are used to check the operation of the integrated circuits being manufactured and are normally employed on fabrication lines for quality control and monitoring during fabrication. The SEM has, indeed, underpinned the progress made in the semiconductor industry over the years and is most likely to continue to play such a role in the future.

In biological applications, the SEM is equally important. However, because the samples are of tissues from living beings and/or plants, these are normally 'specially' treated to withstand the low pressure of the SEM (e.g. in a process called dry-freezing, where the sample is first subjected to a low temperature treatment to freeze-dry it followed by slicing and inspection). This is normally followed by coating its surface with a conducting film to avoid being charged by the incident electrons. These special treatments have by and large been avoided in the ESEM, which is now becoming an important imaging tool in biological applications. But, on the other hand, there are still a lot of instruments which are not of that type in addition to there being some applications which may require inspections in conventional SEMs. In these cases the sample would normally be covered by a conducting metal film prior to inspections. Such films are so thin that they still allow one to observe fine details.

6.7 Fourth-Dimension Electron Microscopy or Time-Resolved Electron Microscopy

In 2009, the Arab laureate, Ahmed Zewail, announced the success of his research group in introducing a new dimension to electron microscopy by utilising femto-second laser pulses to radiate the electron cathode of the microscope, thus releasing single packets of electrons for bombarding the samples instead of a continuous beam. The laser used is carefully chosen to just exceed the work function of the microscope's electron source, which is normally of the field electron emitter type, as discussed in Sect. 6.6.1 above. The bound electrons at the Fermi level of the electron source are released upon the laser pulse striking the tip of the electron emitter. The released 'packet' of electrons is then accelerated by the electron optical column towards a sample under investigation. Note that these electrons are also focussed by the column in exactly the same way as a continuous stream of electrons would have been; so in effect one is using a focussed packet of electrons to bombard the sample. Zewail called this new microscopy mode ultra-fast electron microscopy, which now has come to be known as 4D UEM (see Chap. ▶ 3).

In these experiments, one can follow the dynamic response of the sample at such incredibly short time scales.

The pioneering work of Zewail is in carefully choosing the wavelength of a laser (which corresponds to a given energy) to slightly exceed the work function of the electron source of the microscope. This then eliminates the need to apply a voltage pulse to the emitter, which is limited to microseconds at best. Laser-pulsing at the femtosecond is now standard in 4D electron microscopy. It should be said that 4D electron microscopy is in its infancy, and there are several research groups around the world who are actively engaged in the development of this exciting 4D microscopy technique (see Chap. ▶ 3 for a full list of references and the various areas of application of the technique). Their efforts are likely to open up new areas of applications and enhance our understanding of materials and biological processes.

6.8 Lensless Electron Microscopy

The 1986 Nobel Prize in Physics was awarded for the development of a new type of electron microscopy called scanning-tunnelling microscopy (STM) and for the development of the first electron microscope, given to one of the survivors Ernst Ruska. The STM was demonstrated by two Swiss scientists, Heinrich Rohrer and Gerd Binnig, who proposed a new method to map surfaces at the atomic scale. In this technique, a voltage-biased sharp-field electron emitter, as discussed above, is brought ever so close to the specimen surface (i.e. in the order of only several atomic layer distances) where electron tunnelling via field emission starts with the application of very small voltages of less than 50 V. The emitter is then scanned in a raster fashion over the surface, and the tunnelling current obtained is used to map the spatial distribution of the surface atoms. The original STM experiment was essentially carried out under ultra-high vacuum conditions (i.e. about 10^{-10} mbar). Today there is an array of different configurations of the same principle in what is referred to as scanning-probe microscopy (SPM). It is ironic that the majority of the SPM methods can be carried out under atmospheric pressure. Scanning-probe microscopy today is used in a variety of disciplines ranging from biological applications to physics and engineering. The reader is referred to recent publications for up to date references [23].

Finally, whilst much work has been invested in developing ever more sophisticated electron lenses to map surfaces at the highest possible resolution, a recent development has demonstrated the possibility of producing high resolution secondary electron images at the sub-10 nm region without the use of an electron lens [24]. There is a great deal of research still to be carried out to optimise the technique and to quantify it, but this has been an interesting development in electron microscopy and could yet lead to further if not different areas of applications of this indispensable instrument.

6.9 Application of Electron Microscopy Towards Light-Producing Devices

In this international year of light and light-based technologies, it is appropriate to cite one area where electron microscopy has been an invaluable research tool in its own advancement. The use of electron microscopy has over the years been fundamental in developing the light-producing devices we have been using, ranging from the incandescent light bulb to the latest development of known as ‘light-emitting diodes (LEDs)’ (see later on an explanation of LED). There are a number of benefits in developing an efficient ‘white’ light source like LEDs closely

resembling natural light. For example, it is known that sunlight is responsible for replenishing 90 % of the vitamin D in human bodies (via exposing our skins to direct sunlight), compared to only 10 % replenishment via food [25]. This vitamin is of great medical importance to humans. Natural light is also believed to have a protective effect against certain types of cancers (breast and prostate types) by preventing the over-production of cells [26]. In addition, its deficiency could lead to a weak immune system against colds, coughs, etc., and many studies link it to body fatigue (seasonal affective disorder—SAD), broken bones and fracture. The conventional light bulb as well as the fluorescent tubes is very inefficient in resembling sunlight, which makes the search for alternatives all the more important.

6

In addition to the medical benefits LEDs bring to human life, their development has potential impact on the amount of electricity used worldwide. It is currently estimated (in the year 2015) that lighting amounts to at least 20 % of the total world electricity produced. For a country like the UK, for example, this amounts to the equivalent cost of about $3bn/annum. If white-colour-based LEDs could replace the conventional light bulbs around the house and in offices, then energy saving of more than 80 % could result. The US Department of Energy estimates a projected energy saving amounting to about $30bn/annum, if LED replace incandescent lighting, with the added bonus of avoiding about 1800 million metric tons of carbon emissions [27]. Imagine this being applied worldwide; it would have huge savings on world resources but would particularly allow less developed countries to offer their citizens lighting resources at modest cost. Furthermore, LEDs are more compact in size, contain no mercury—hence safer to use and dispose of—and have more than 30 times longer life (~50,000 h of use).

But what is a light-emitting diode, and what does it consist of? And more crucially what does electron microscopy have to do with its development?

The light-emitting diode is one of many types of the basic diode that are used as the building blocks in semiconductor devices. However, there are some crucial differences between diodes used in general electronic devices, such as the indispensable silicon p-n junction diodes used in most electronic devices, and those employed for lighting. In the former case, a piece of silicon, for example, is designated into two parts. In one part a material like boron (B) replaces the host silicon atoms to a value not exceeding 0.01 % of the total volume which makes the p-region, while on the adjoining part a material like phosphorous (P) is used to a similarly low concentration and this makes the n-region of what becomes a p-n device. The n-region is the one which has more electrons per unit volume than the host silicon whilst the p-region has more 'holes', i.e. less electrons. It is the electron movement within the two regions which causes the flow of electrical signal throughout the electron devices.

The development of white LED technology has eluded scientists for many years. Red and green LEDs have been available for almost 50 years, but it was the blue LED that was required to make up the white light. The case for a successful and functioning blue LED is rather special, though. For brevity, one can summarise the material requirements for producing blue LEDs to be:

1. The semiconductor material has to be of direct type (i.e. when an electron loses energy by falling across the band gap, no phonons—which leads to heat—are also produced.). For more explanation of direct and indirect band gap semiconductors, see Chap. ► 10.
2. The energy gap should be intermediate in the range 1.77 and 3.1 eV, which corresponds to a wavelength in the visible spectrum between 400 and 700 nm.
3. The material should be amenable to the formation of p-n junction diodes.

The above requirements eliminate widely used semiconductor materials like silicon and germanium which are both of the indirect type (see [28] for more

discussion of semiconductor fundamentals). Although this is the case, this did not stop researchers from developing LEDs based on silicon compounds, such as silicon carbide (SiC) which has been successfully used to produce commercial LEDs, albeit with very low efficiency.

The search for a material or a compound of materials to satisfy all the above three requirements looks as if it has finally been found in gallium nitride (GaN). This material is a direct III–V semiconductor with an energy gap of 3.36 eV. It is now considered as one of the leading candidates in this technology and commercial devices using this material have already been on sale for the last year or two. The development of this compound material owes heavily to the work and contribution of two groups, *Isamu Akasaki*, *Hiroshi Amano* of Nagoya University, Japan and *Shuji Nakamura* of the University of Santa Barba Ca, USA, who shared the Nobel Prize for Physics for 2014 'for the invention of efficient blue light-emitting diodes which has enabled bright and energy-saving white light sources'. Whilst they demonstrated the blue-light laser using GaN during the late 1980s and early 1990s, it took many research groups in academia and industry alike all over the world nearly three decades to develop efficient methods for the reliable and large-scale production of GaN. Professor Sir Colin Humphreys of the University of Cambridge, UK (see Chap. ▶ 5) has been one of the world's leading scientists in developing GaN-based LEDs, and his research group's work has built upon the demonstration of the Japanese Nobel Prize winners and developed new methods for the large-scale production of commercial material. The use of electron microscopy has been crucial in this endeavour, and examples from work published by Humphreys' group will be briefly reviewed here, although parallel stories indicative of the crucial role that electron microscopy has played apply in the case of all the other examples and in the various developments by scientists all over the world towards this goal.

The challenge faced by the community researching GaN for use in LED technology, following the Japanese demonstration, was in producing high-quality material free of defects. It is important to appreciate that the role of defects in semiconductors is similar to that of the 'dopant' materials one adds to form the p-n junctions. Further, the small amount of this foreign material or defect that alters the electrical behaviour of the semiconductor is comparable in size to that of a golf ball and a football pitch. However, in terms of the foreign material one adds to make a p- or n-type semiconductor, one knows a priori its characteristics and how much to add to achieve the required results. In the case of material defects, this is uncontrolled with devastating consequences, if their numbers exceeds a certain limit.

One of the challenges faced by Humphreys' group was in using silicon wafers as the base on which to grow the various layers of materials needed to produce the LED device. A schematic of a typical LED device based on GaN is shown in ◘ Fig. 6.17 [26]. The substrates used in this technology and almost all other semiconductor devices, and which do not have an active role in the working of the device, are normally made from silicon wafers to reduce the cost of the more expensive active materials like GaN. However, it is the substrate material that causes some of the dislocation defects seen in the grown GaN materials. Such dislocations reduce the efficiency of the LED and their reduction is therefore of utmost importance in the search for efficient devices.

In addition, depositing GaN on Si causes the former to react with the Si to form a Ga-Si alloy and melt-back etching that alters the concentration of the constituting elements. To avoid this from happening, a layer of aluminium nitride (AlN) is first deposited on the Si as a nucleation layer. The quality of this layer and its interface with the Si substrate are of great importance for the produced LEDs; and the interfacial layer between the Si and the AlN, as depicted in ◘ Fig. 6.18, was only possible to study by using a state-of-the-art TEM with a resolution of 0.1 nm.

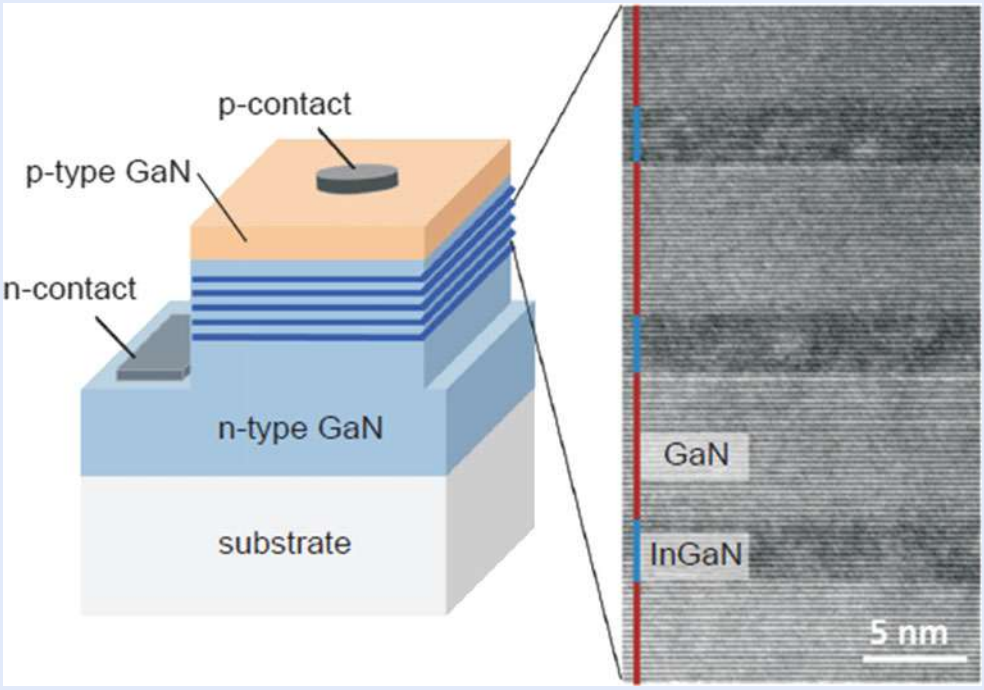


Fig. 6.17 A schematic of an InGaN/GaN quantum well LED (courtesy of Prof. Humphreys) [29]

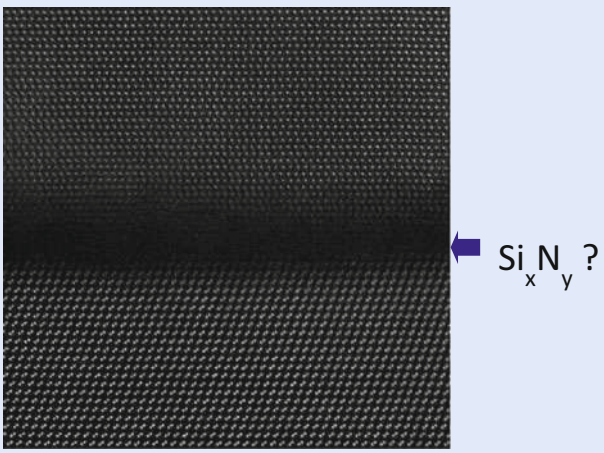


Fig. 6.18 TEM Image from—Cs corrected instrument showing the lack of crystallinity at the interface between the substrate and the deposited semiconductor (courtesy of Prof. Humphreys) [29]

This TEM resolution could only be achieved after the correction of the microscope's spherical aberration. It is interesting to note that the solution offered by Humphreys' research group has now successfully been taken up by UK industry (Plessey Semiconductors).

It is also interesting to note that whilst most researchers thought that the use of indirect bandgap semiconductors is not suitable for efficient LEDs, nanotechnology is providing a route where this may not necessarily be the case. The work reviewed by Professor Nayfeh in this book demonstrates that indeed there is life for silicon in producing efficient lighting if one uses silicon nano-particles (see Chap. ► 10).

6.10 Conclusions

The development of the blue LED, which earnt its developers the Nobel Prize in 2014, and the subsequent tireless work of scientists and engineers in its development as a working device, could not have been realised had they not used state-of-the-art TEMs which are equipped with aberration correctors, but equally if scientists and engineers had not followed the scientific method of enquiry in their research and development. The use of the law of refraction is fundamental in electron and light optics, and the name of Ibn Sahl, the tenth century Arab mathematician, should be on a par with that of Snell. It is also clear that the correction of the spherical aberration defect in electron optical instruments and in other imaging devices is crucial for achieving the ultimate performance of such devices. Whilst Ibn al-Haytham's early discovery of spherical aberration should be correctly acknowledged, as well as his many other pioneering work in optics, his legacy in science goes far beyond these achievements (see Chap. ▶ 2 for more details). More importantly in relation to science and engineering, Ibn al-Haytham again deserves credit and recognition for his role in laying down the foundations of the scientific method of enquiry in his work on optics some 1000 years ago entitled, Kitab al-Manazir. It is fitting therefore for UNESCO to recognise this year and celebrate the contribution of Ibn al-Haytham as a pioneering polymath.

References

1. Rashed R (2005) Geometry and dioptrics in classical Islam. al-Furqan Islamic Heritage Foundation, London
2. Dahl PF (1997) The flash of the cathode rays. CRC, Boca Raton
3. Leicester HM (1971) The historical background of chemistry. Dover Publications Inc, New York
4. Grivet P (1972) Electron optics, 2nd English edn. Oxford, Pergamon.
5. Prutton M, El-Gomati MM (2004) Scaning Auger Electron Microscopy, John Wiley & Sons, Chichester
6. Fultz B, Howe J (2013) Transmission electron microscopy and diffractometry of materials. In: Di Meglio J-M, Rhodes WT, Scott S, Stutzmann M, Wipf A (eds) Graduate texts in physics. Springer, Heidelberg
7. Chescoe D, Goodhew PJ (1984) The operation of the transmission electron microscope. Royal Microscopical Society Microscopy Handbooks 02, Oxford University Press, Oxford
8. Orloff J (ed) (2008) Handbook of charged particle optics, 2nd edn. CRC, Baton Rouge
9. Bronsgeest M (2014) Physics of Schottky electron sources. Pan Stanford Publishing, Pte, Singapore
10. Charbonnier F (1990) Developing and using the field emitter as a high intensity electron source. Appl Surf Sci 94–95:26–43

11. Milne WI, Teo KBK, Mann M, Bu IYY, Amaratunga GAJ, De Jonge N, Allioux M, Oostveen JT, Legagneux P, Minoux E, Gangloff L, Hudanski L, Schnell J-P, Dieumegard LD, Peauger F, Wells T, El-Gomati M (2006) Carbon nanotubes as electron sources. Phys Status Solidi 203 (6):1058–1063
12. Wahab H (2012) Metal oxide catalysts for carbon nanotubes growth: The growth mechanism using NiO and doped ZnO. PhD thesis, University of York, UK, York
13. Algarni, Hamed (2013) Synthesis and Characterizations of ZnO Nanostructures for Field Emission Devices. PhD thesis, University of York, UK, York
14. de Jonge N, Lamy Y, Schoots K, Oosterkamp TH (2002) High brightness electron beam from a multi-walled carbon nanotube. Nature 420:393–395
15. Krvanek OJ, Delby N, Murfitt M (2009) Handbook of charged particle optics, 2nd edn. CRC, Boca, Boca Raton
16. Yu Z, Pan Y, Shen Y, Wang Z, Ong Z-Y, Xu T, Xin R, Pan L, Wang B, Sun L, Wang J, Zhang G, Wei Zhang Y, Shi Y, Wang X (2014) Towards intrinsic charge transport in monolayer molybdenum disulfide by defect and interface engineering. Nat Commun 5, 5290
17. The Hubble telescope mission web site. ► https://www.nasa.gov/mission_pages/hubble/main/index.html
18. El-Gomati MM, Zaggout F, Jayakody H, Tear S, Wilson K (2005) Why is it possible to detect doped regions of semiconductors in low voltage SEM: a review and update. Surf Interface Anal 37:901–911
19. Walker CGH, Zaggout F, El-Gomati MM (2008) The role of oxygen in secondary electron contrast in doped semiconductors using low voltage scanning electron microscopy. J Appl Phys 104:123713
20. El-Gomati MM, Mullerova I, Frank L (1998) The electron. IOM communications Ltd, London, pp 326–333
21. Spallas JP, Silver CS, Murray LP, Wells T, El-Gomati MM (2006) A manufacturable miniature electron beam column. Microelectron Eng 83(4–9): 984–985
22. Goldstein J, Newbury DE, Joy DC, Lyman CE, Echlin P, Lifshin E, Sawyer L, Michael JR (2003) Scanning electron microscopy and X-ray microanalysis, 3rd edn. Springer, Heidelberg
23. Birdi KS (2003) Scanning probe microscopes, applications in science and technology. CRC, Boca Raton
24. Kirk TL (2010) Near field emission scanning electron microscopy. PhD thesis, Swiss Federal Institute of Technology, Zürich
25. Dowed J (2012) The vitamin D cure. John Wiley & Sons, Hoboken, New Jersey
26. SACN Update on Vitamin D (2007) Public Health England report, TSO, London
27. USA Department of Energy report (2014) Office of communication, Washington
28. Sze SM (1981) Physics of semiconductor devices, 2nd edn. Wiley, New York
29. Humphreys, see chapter 5 this book.

Applications

Contents

The Dawn of Quantum Biophotonics

Dmitri V. Voronine, Narangerel Altangerel, Edward S. Fry, Olga Kocharovskaya, Alexei V. Sokolov, Vladislav V. Yakovlev, Aleksey Zheltikov, and Marlan O. Scully

D.V. Voronine (✉) • A.V. Sokolov • M.O. Scully
Texas A&M University, College Station, TX 77843, USA

Baylor University, Waco, TX 76798, USA
e-mail: scully@tamu.edu

N. Altangerel • E.S. Fry • O. Kocharovskaya • V.V. Yakovlev
Texas A&M University, College Station, TX 77843, USA

A. Zheltikov
Texas A&M University, College Station, TX 77843, USA

Moscow State University, Moscow, Russia

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_7

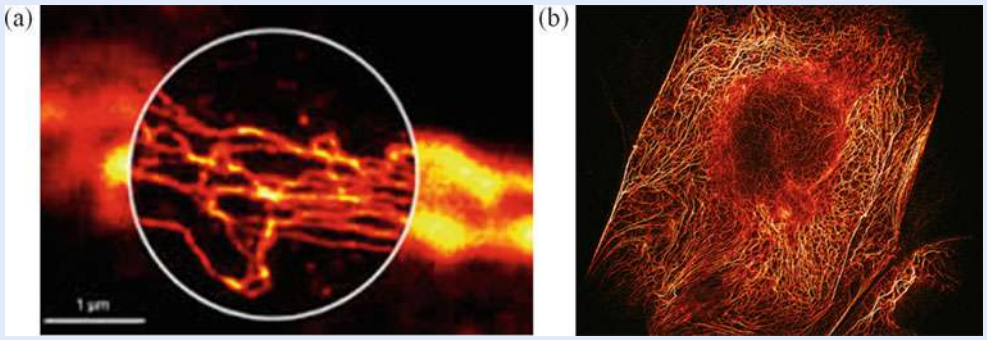


Fig. 7.1 (**a**) STimulated Emission Depletion (STED) microscopy with enhanced resolution (inside the circle) compared with the conventional optical microscopy (outside of the circle). (**b**) REversible Saturable OpticaL Fluorescence Transitions (RESOLFT) image of keratin in cells. Adapted from [1, 2]

7.1 Overview: Toward Quantum Agri-Biophotonics

Quantum mechanics, the crowning achievement of twentieth century physics, is yielding twenty-first century fruit in the life sciences. For example, the broad and multi-faceted field of *Quantum Biophotonics* is exciting and rapidly developing, with many emerging techniques and applications. The broad range of topics includes remote sensing with applications toward plant phenotyping, single cell/virus/biomolecule detection, and superresolution imaging. Progress in the latter was recognized with the 2014 Nobel Prize in Chemistry (Fig. 7.1). New developments in this field are promising and may deliver at least an order of magnitude improvement.

Laser spectroscopy has been widely used for chemical analysis of the living systems. For example, Raman and infrared (IR) spectroscopies can probe the vibrational states of molecules in order to determine, e.g., chemical assay and temperature profile. IR spectroscopy is widely used because it is simple and inexpensive. Raman spectroscopy is more complicated but has many advantages and is a more versatile and powerful tool. Coherent and stimulated Raman techniques can be used to increase the speed and strength of signal acquisition by orders of magnitude. Femtosecond adaptive spectroscopic techniques for coherent anti-Stokes Raman spectroscopy (FAST CARS) was used to detect small amounts of anthrax-type endospores on a nanosecond time scale [3–5] and inside a closed envelope [6]. Fluorescence measurements of fecal matter in water achieved increase of sensitivity by three orders of magnitude [7]. Laser-induced breakdown spectroscopy has been developed to perform studies of plant physiology and phenotyping. Surface-enhanced coherent Raman spectroscopy achieved astonishing results in detection sensitivity [8, 9]. All these techniques are aimed at increasing the speed and reliability of field-based sensing and can be used for improving crop yield. These scientific and technical innovations hold promise for new diagnostic tools bridging the gap between fundamental quantum research and potential agricultural applications.

7.2 Fundamental Light–Matter Interactions and Spectroscopy of Biological Systems

Quantum mechanics provides the most complete description of the light–matter interactions and of the various spectroscopic techniques used for probing the structure–function relations in many fields of science and engineering. These

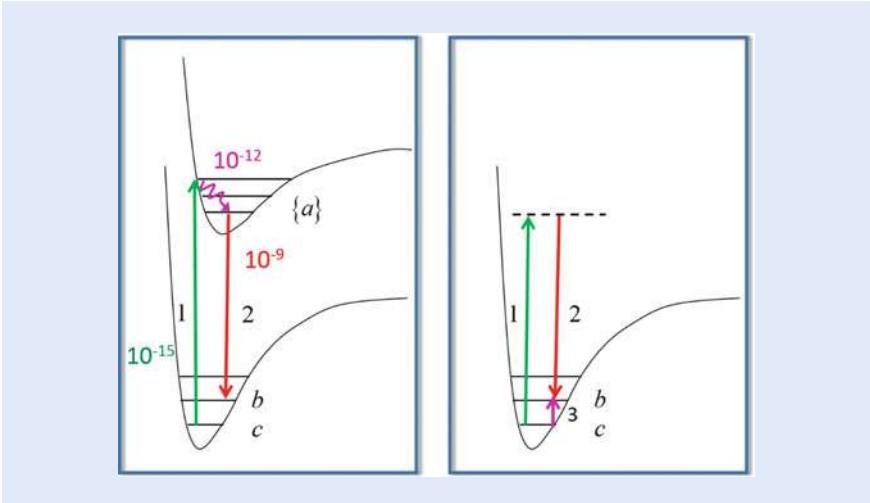


Fig. 7.2 Fundamental light–matter interactions in bio-objects involve absorption, emission, and scattering of light (represented by *arrows*) by molecular systems (represented by *energy level diagrams*). Absorption, relaxation, and fluorescence emission take place on a time scale of 10^{-15} s, 10^{-12} s, and 10^{-9} s, respectively (*left panel*). Non-resonant Raman scattering occurring through a virtual level (*dashed line*) quasi-instantaneously (*right panel*). Infrared absorption (*purple arrow*, *right panel*) provides complimentary information to Raman spectroscopy

interactions happen on different time scales from femtoseconds [~10^{-15} as in the case of absorption shown by a green arrow 1 in Fig. 7.2 (left)] to picoseconds [~10^{-12} as in the case of vibrational cooling shown by a purple wavy arrow in Fig. 7.2 (left)] to nanoseconds [~10^{-9} as in the case of fluorescence emission shown by a red arrow 2 in Fig. 7.2 (left)] and longer. Some processes are almost instantaneous, for example, Raman scattering [shown by red arrow 2 in Fig. 7.2 (right)]. Modern pulsed laser sources with pulse durations from nanoseconds to femtoseconds and beyond can be used to investigate and control these processes. Fluorescence and Raman scattering are major laser spectroscopic techniques which can be used for analytical purposes in agricultural, biomedical, environmental, and many other applications. Infrared absorption also provides complimentary information. Below we describe several examples of recent breakthroughs using these and other techniques within the general field of quantum biophotonics.

The Raman spectroscopic technique is a valuable tool which has excellent potential for the analysis of plant and animal agriculture. Raman spectroscopy is a vibrational measurement which can be applied directly to plant and animal tissues yielding characteristic key bands of individual constituent components. The physics of the Raman effect is analogous to scattering off of an oscillating mirror. It relies on inelastic scattering of light from a laser in the visible, near-infrared, or near ultraviolet range off of molecules. The laser light interacts with molecular vibrations resulting in the energy of the laser photons being down shifted (Stokes) or up shifted (anti-Stokes) signals, respectively (Fig. 7.3). The shift in energy gives useful information about the vibrational modes in the system providing molecular "fingerprints" which can be used for identification. Moreover, essentially all molecules have Raman active vibrational transitions.

Unlike optical absorption transitions, it is possible in principle to separate out the Raman emission of a particular target molecule from the many background molecules in the focal volume. This is because the molecular vibrational levels scatter light with distinctive frequency shifts that are often narrowband. Since there are usually many vibrational transitions in a biomolecule, there are many Raman lines to choose from and these can be used to fingerprint the target molecule. Another important advantage of Raman techniques is that the laser

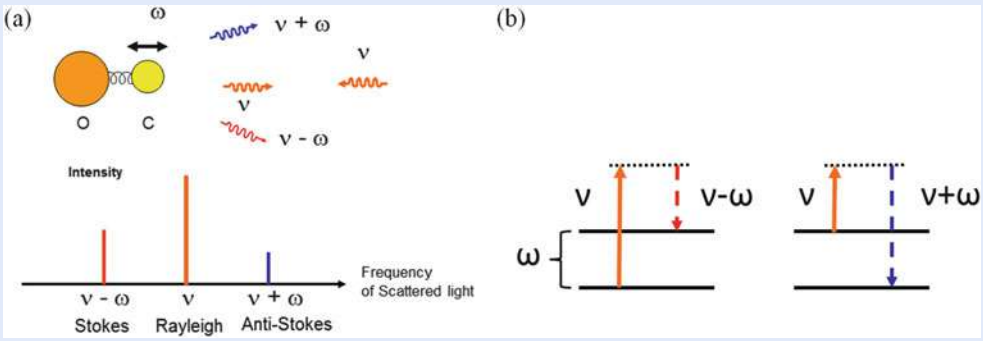


Fig. 7.3 The Raman effect (**a**) and the energy level Stokes and anti-Stokes diagrams (**b**) for a simple molecule. Light at the probe frequency *v* is inelastically scattered off of the molecule resulting in the characteristic frequency shift which serves as a fingerprint for chemical analysis

does not need to be tuned near the optical transition. Even if the optical transition is in the UV, Raman transitions can be efficiently excited with lasers tuned as far away as the near IR, for example, 1064 nm or longer. For such large detunings, the excited state is essentially unpopulated, and so laser damage effects (such as bleaching) or background from chlorophyll fluorescence is strongly suppressed. These features make the Raman approach particularly useful for plant and animal studies which deal with highly complex samples and their environment. Figure 7.4 shows an example of application of Raman spectroscopy for the early detection of anthocyanin markers of stress in plants [10].

Brillouin scattering is another light–matter interaction phenomenon which has delivered an emerging biomedical tool that has already been used to study bone, collagen fibers, cornea, and crystalline lens tissue. Unlike Raman spectroscopy, which offers information about the chemical makeup of the sample, Brillouin spectroscopy provides information about the viscoelastic properties of a material, and consequently, can characterize larger bulk changes. Each of these imaging tools offers useful diagnostic information. Therefore a single apparatus that could provide simultaneous measurement of both spectra from the same point would be extremely powerful for sample characterization and analysis [11]. However, if separate instruments are used, the acquisition time for each Brillouin spectrum is very long (~15 min), making such approach impractical. The lack of the same-point detection for both spectra makes the analysis complicated. To overcome these issues, we recently used a single pump laser to generate both Raman and Brillouin spectra [12] and to provide simultaneous imaging from the selected confocal volume. More importantly, we take advantage of the recent advancements in Brillouin spectroscopy to decrease the acquisition time, as any practical implementation of the simultaneous detection requires that the times for both to be comparable. Unlike other approaches that use scanning Fabry Perot cavities, we utilized a virtually imaged phase array (VIPA), which offers a higher throughput efficiency, >80 %, and does not require scanning to extract a complete spectrum. Subsequently, a VIPA-based system drastically cuts down the acquisition time, which was traditionally a limiting factor in Brillouin spectroscopy. The major challenge in Brillouin spectroscopy of biological systems is eliminating the large amount of elastic scattering, which makes it difficult to identify a weak Brillouin peak. We recently reported that this limitation can be overcome by using a molecular/atomic gas cell as a notch filter [13]. Utilizing these advancements, we demonstrate simultaneous Raman–Brillouin microscopy, a potent new tool for bioimaging and analytical characterization.

Both Raman and Brillouin phenomena arise from the inelastic scattering of light, where the scattering causes the frequency of light to shift in accordance with

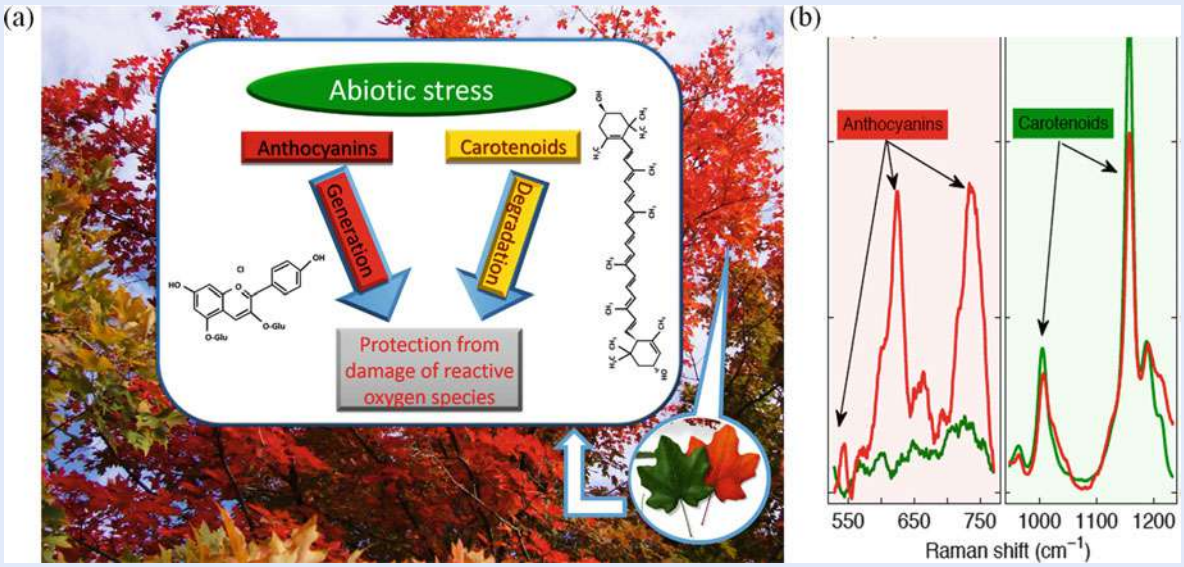


Fig. 7.4 (**a**) The autumn colors of maple trees are due to anthocyanins and are indicative of plant stress. To enhance the survival probability of stressed crops, it is important to identify stressed plants as early as possible to allow for effective intervention. Raman spectroscopy can reveal early spectroscopic signatures of carotenoids and anthocyanins. (**b**) Raman spectra of the unstressed (*green*) and saline stressed (*red*) plants after 48 h. Adapted from [10]

some resonant property. In the case of Raman, the incident light interacts with molecular vibrations causing the light to shift. Similarly, Brillouin scattering is caused by the inelastic interaction of light with periodic fluctuations in a material's index of refraction. From the quantum physics perspective, Brillouin scattering is an interaction between an electromagnetic wave and a density wave (photon–phonon scattering). Thermal motions of atoms in a material create acoustic vibrations, which lead to density variations and scattering of the incident light. These fluctuations carry information about a material's bulk compressibility and viscoelasticity. Whereas Raman scattering can have frequency shifts on the order of 100 THz, Brillouin shifts are only on the order of 10 GHz due to the relatively low energy of the acoustic phonons. The magnitude of the Brillouin shift is dependent upon the collection geometry. Through the measured Brillouin shift we are able to extract mechanical properties of the material, including the speed of sound, adiabatic compressibility, and the longitudinal modulus. While Brillouin scattering is most often used to measure elastic properties of materials, the linewidth of the Brillouin signal also provides information about the viscoelasticity.

The imaging capability of the Raman–Brillouin microscope is shown using a T-shaped sample made of two materials with different mechanical and chemical properties (Fig. 7.5). Cyclohexane and poly(ethylene glycol) diacrylate (PEGDA) hydrogel are model systems, and the hydrogel was cured in a "T"-shaped mold to provide spatial contrast. The body of the T-shaped structure was created using PEGDA. The T-shaped structure was then placed into a solution of cyclohexane, which provided contrast for the image. The corresponding Raman and Brillouin images are shown in Fig. 7.5. In all cases a high-contrast image was produced, whose spatial accuracy can be confirmed when compared to an optical photograph of the sample.

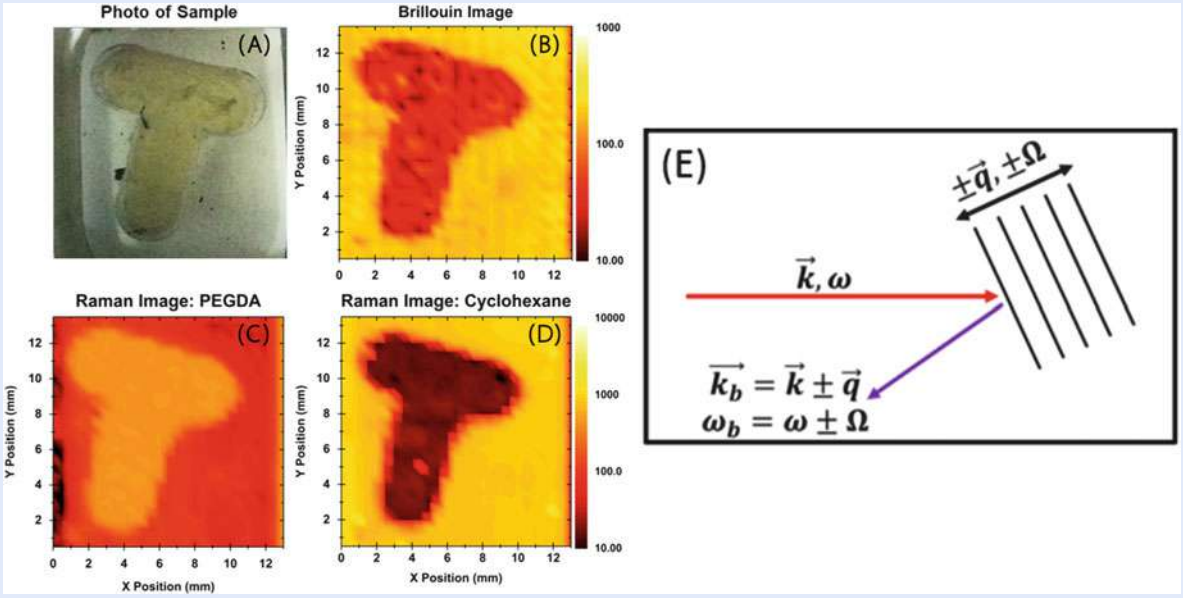


Fig. 7.5 Images of the T-shaped sample: (**a**) Optical photograph; (**b**) Brillouin image using the intensity of the cyclohexane peak at 4.58 GHz as contrast; (**c**) Raman image using the normalized ratio between the PEGDA peaks at ~1400 and ~2800 cm^{-1}; (**d**) Raman image using the cyclohexane peak near 800 cm^{-1} as contrast. (**e**) Schematic diagram of the Brillouin scattering process where incident light interacts with the acoustic field of the material. The magnitude of the frequency shift is dependent upon the direction of light scattered by the acoustic wave. Adapted from [12]

7.3 Quantum-Enhanced Remote Sensing

7.3.1 Anthrax Detection in Real Time

Chemically specific optical imaging and sensing techniques play a pivotal role in developing preventive measures to maintain chemical and biological safety. Frequently, objects to be imaged and identified are not present on a surface, and the light probe has to pass through layers of scattering material. For example, the anthrax-causing bacteria *Bacillus anthracis* can form micron-sized spores which can be present in air, or on the skin of cattle. Brucellosis-causing bacteria *Brucella* spp. can be present in strongly scattering liquids such as milk or blood.

In particular, the detection of anthrax in the farmland environment is an important problem of current interest. Clearly, it would be highly desirable to identify the chemical content of, for instance, air near farm animals, or in wool, and so forth, i.e., identify anthrax spores on-the-fly in air, or image samples of hair in order to chemically distinguish common biological and chemical threats. However, the detection of a chemically specific target in a harsh environment presents several problems. First, targeted molecules are not directly accessible for interrogation, and a remote sensing technique has to be used. On passing through the scattering medium (e.g., wool), both the incident and signal light are substantially diminished, weakening the signal and diffusing the spatial information about the signal's origin. Second, the chemical specificity of anthrax detection has to be maintained in the presence of a substantial background from surrounding chemicals.

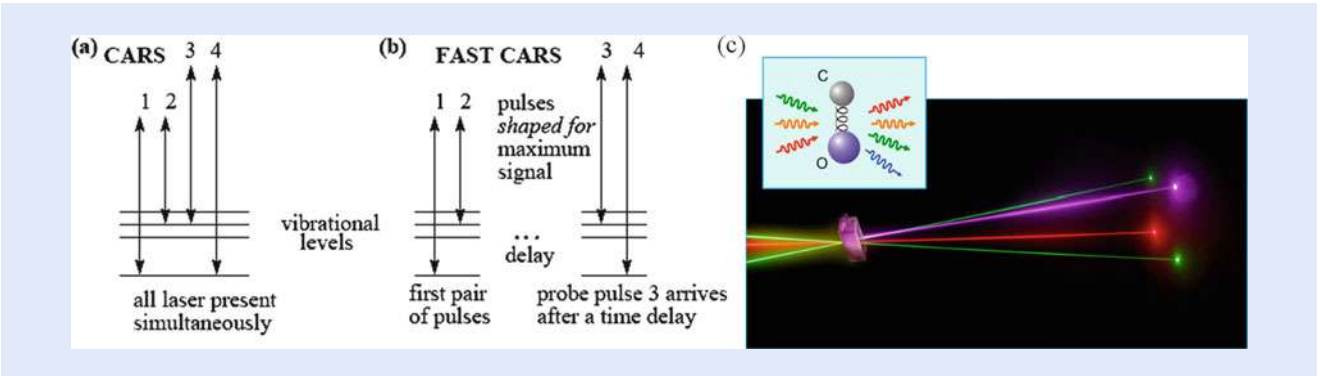


Fig. 7.6 Femtosecond adaptive spectroscopic techniques for coherent anti-Stokes Raman spectroscopy (FAST CARS). Comparison between the energy level schemes and laser configurations for the conventional CARS (**a**) and FAST CARS (**b**). (**c**) Sketch of the laser beam configurations. Inset shows a vibrating molecule emitting signals

7

The major drawback of the commonly used spontaneous Raman spectroscopy is the weak signal strength. We have developed a new spectroscopic technique (FAST CARS) based on maximizing quantum coherence by breaking the adiabaticity** and laser pulse shaping to optimize it. In FAST CARS spectroscopy the signal is proportional to the number of molecules squared, as opposed to the linear dependence in the spontaneous Raman technique [14]. In addition, FAST CARS was the first proposal for background suppression in precision sensing of minor molecular species within a highly scattering environment [3], and provided an efficient solution to the problem of detecting and identifying anthrax-type bacterial endospores in real time [4, 5].

Figure 7.6 shows the basic idea behind the FAST CARS approach. With standard CARS (Fig. 7.6a) two lasers with frequencies ω_1 and ω_2 (we often refer to these lasers as 1 and 2) are incident on a sample. The difference frequency may be resonant with some molecular vibrational excitation. At the same time, the sample may also absorb a third photon (either a third laser with frequency ω_3, or a second photon from laser 1, $\omega_1 = \omega_3$) and generate light at frequency $\omega_4 = \omega_1 - \omega_2 + \omega_3$. Variations of this scheme are numerous. For example, one of the two lasers 1 or 2 might be resonant with an electronic excitation in the molecule.

The problem with these schemes is that the process is masked by four-wave mixing (FWM) in a non-resonant medium. This produces broadband nonlinear generation that can be much larger than the small, vibrationally resonant CARS process. That is, the FWM generation at the detected wavelength range can be much larger than the signal we want to detect that is resonant with a specific molecular vibration. Various techniques have been proposed for suppressing this non-resonant process, including heterodyne detection of the CARS signal and the use of polarization tricks to suppress the undesired signals. FAST-CARS was developed for suppressing the nonresonant background based on pulse shaping and is shown in Fig. 7.6b. Here, all lasers provide short pulses. Pulses 1 and 2 are applied to the sample first and laser pulse 3 is delayed. When laser number 3 is applied to the sample and lasers 1 and 2 are not, non-resonant processes cannot occur. However, beams 1 and 2 will have excited coherence between the vibrational levels in the molecule for which they are Raman resonant. If this coherence last longer than the delay, laser 3 will scatter from the coherence and still produce a

**Two-photon resonant pulses produce $\rho_{bc} \neq 0$ quantum coherence by breaking adiabaticity of the molecular excitation; but off-resonant pulses return the molecule to the ground state as $\rho_{bc} = 0$

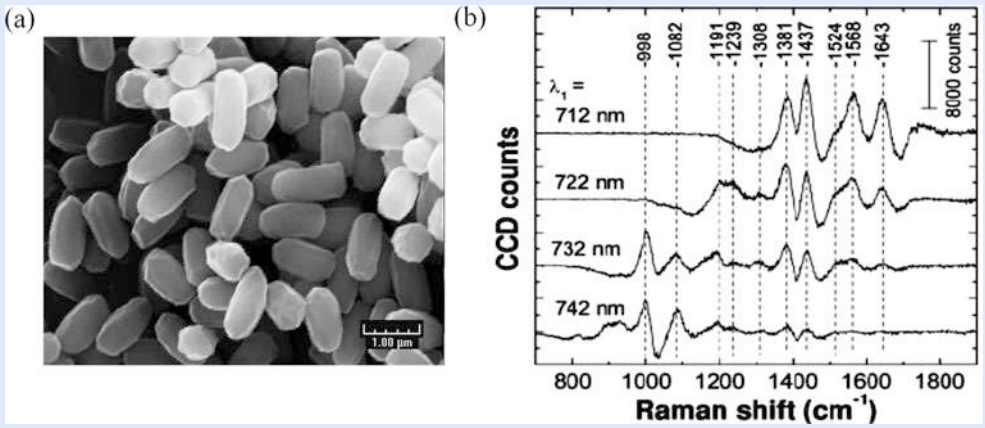


Fig. 7.7 FAST CARS on *Bacillus subtilis* spores (**a**) at various pump wavelengths (λ_1) from 712 to 742 nm (**b**). Adapted from [4]

CARS signal at ω_4. The amount of CARS generation as a function of delay time will show a temporal modulation characteristic of the vibrational energy level structure.

FAST CARS detection of micrometer size *Bacillus subtilis* spores has been performed (Fig. 7.7). A combination of ultrafast pump-Stokes Raman excitation and narrow-band probing of the molecular vibrations provides a species-specific signal from spores [4, 5]. In this scheme, the use of spectrally broad preparation pulses leads to the excitation of multiple Raman lines. The narrow-band probing, on the other hand, allows for frequency-resolved acquisition, as in traditional spontaneous Raman measurements. Recording of the whole CARS spectrum at once makes the technique relatively insensitive to fluctuations. The spectral contrast between the broadband preparation and narrow-band probing provides a way to differentiate between the Raman-resonant and non-resonant contributions. It also helps to mitigate the strength of the non-resonant FWM.

Frequently, objects to be imaged and identified are not present on a surface, and the light probe has to pass through layers of scattering material. In particular, the detection of weaponized anthrax in the mail room is an important problem of current interest. It would be highly desirable to identify the chemical content of an unopened envelope (i.e., image through layers of paper and chemically distinguish common biological and chemical threats from nonhazardous materials). However, the detection of a chemically specific target in a scattering medium presents several problems. First, targeted molecules are not directly accessible for interrogation, and a remote sensing technique has to be used. On passing through the scattering medium, both the incident and signal light are substantially diminished, weakening the signal and diffusing the spatial information about the signal's origin. Second, the chemical specificity of anthrax detection has to be maintained in the presence of a substantial background from surrounding chemicals, such as paper. Recently we demonstrated detection of anthrax-like spores inside a closed envelope using coherent Raman microscopy (Fig. 7.8) [6].

7.3.2 Stand-Off Spectroscopy

The need for an improved approach and efficient tools for remote optical sensing is high since they would facilitate applications ranging from environmental diagnostics and probing to chemical surveillance and biohazard detection. Present-day techniques rely on collecting incoherently scattered laser light and are often hindered by small signal collection efficiency. Availability of a laser-like

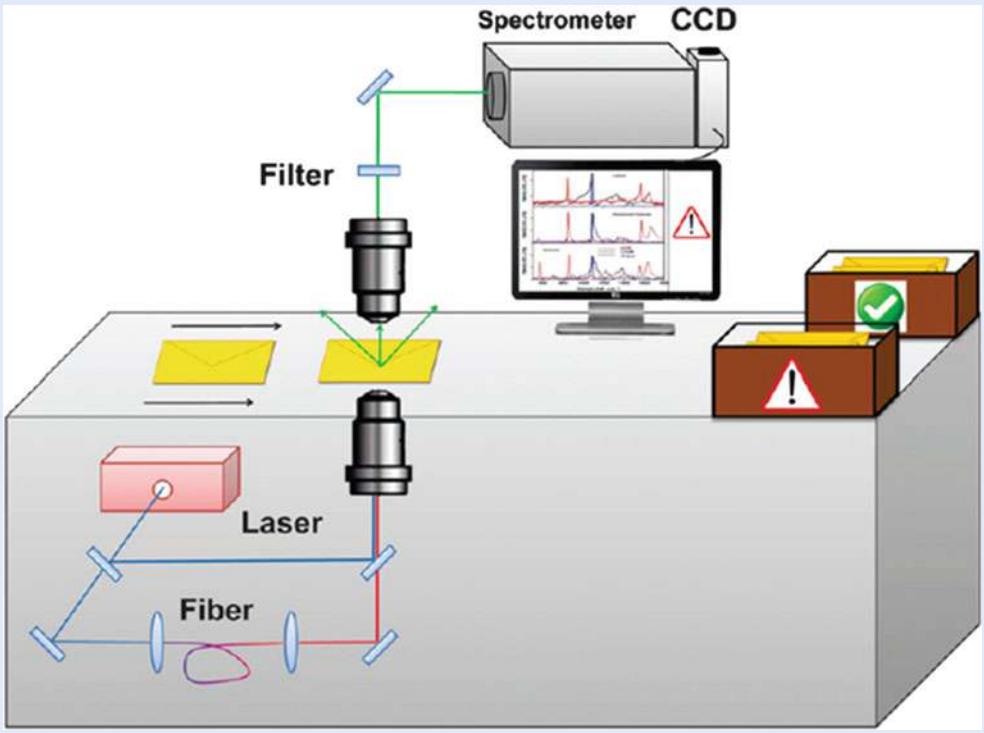


Fig. 7.8 Schematic diagram of the experimental setup for coherent Raman microspectroscopy imaging of anthrax-like spores inside a closed envelope. Adapted from [6]

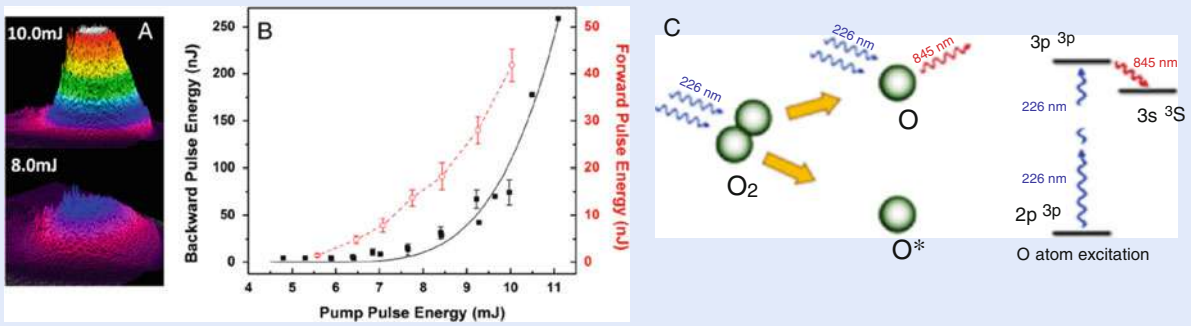


Fig. 7.9 (**a**) Spatial beam profiles of the 845 nm emitted backward pulse at pump energies above (10.0 mJ) and at (8.0 mJ) threshold. (**b**) The energy per pulse of both the forward (*red circles*) and backward (*black squares*) signals versus the pump power. (**c**) Two-photon dissociation of the oxygen molecule and subsequent two-photon resonant excitation of the ground-state oxygen atom fragment result in emission at 845 nm. Adapted from [16, 17]

light source emitting radiation in a controlled directional fashion [15], from a point in the sky back toward a detector, would revolutionize the area of remote sensing. As one example, we have demonstrated the possibility of remote lasing of atmospheric oxygen by using picosecond and nanosecond UV laser pulses that produce atomic oxygen via 226 nm UV pump light, and achieved a bright near-infrared (NIR) laser source at 845 nm wavelength (Fig. 7.9) [16, 17]. Nearly two decades after the work on stimulated emission (SE) performed in the context of flame and flow diagnostics, a renewed interest in laser-like emission from open air is motivated by the need for chemically selective stand-off detection of trace gases in the atmosphere [15].

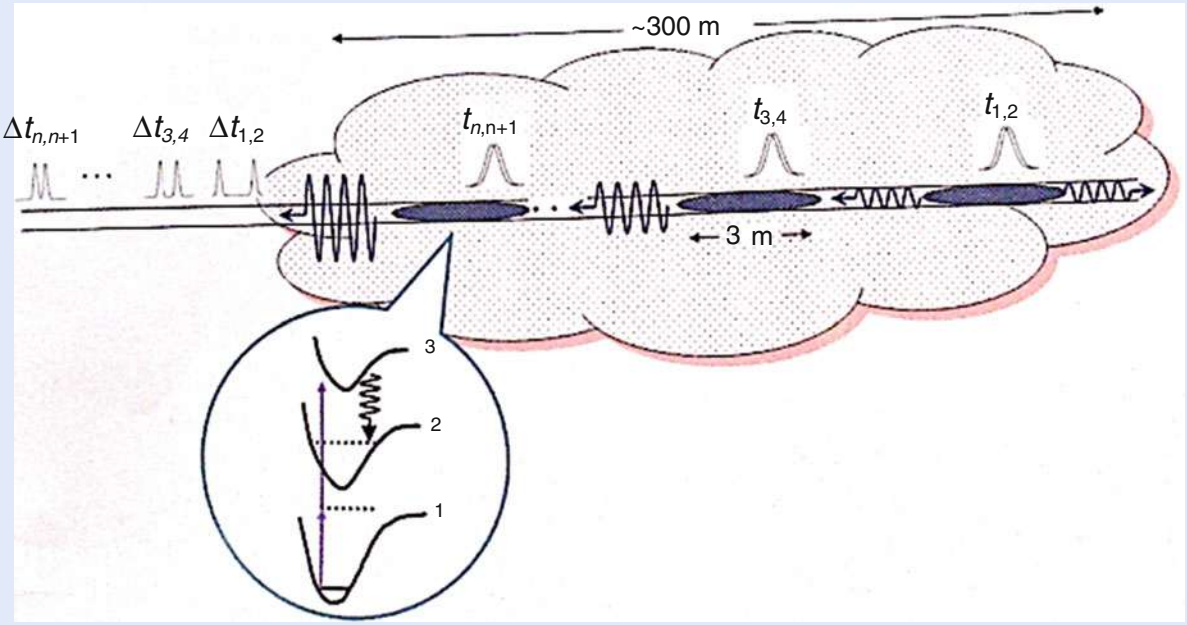


Fig. 7.10 SOS I. Pairs of laser pulses of different colors (e.g., *red* and *blue*) excite a dilute ensemble of molecules in a cloud such that lasing and/or gain-swept superradiance is realized in a direction back toward the observer [18]

Laser-like emission provides a promising tool for a broad class of all-optical stand-off detection methods, as it suggests a physical mechanism whereby a high-brightness, highly directional back-propagating light beam can be generated directly in ambient air. Superradiance can be used to enhance backward-directed lasing in air, using the most dominant constituents such as nitrogen or oxygen. We performed investigations of both the forward and backward-directed emission of oxygen when pumped by nanosecond UV laser pulses. The backward 845 nm beam profile is shown in Fig. 7.9a using nanosecond pulses approximately 10 mJ/pulse of 226 nm. High-quality, strong coherence-brightened emission was observed (Fig. 7.9b).

Earlier, we proposed to use stand-off spectroscopy (SOS) techniques for detecting harmful impurities in air using gain-swept superradiance [18]. In our first SOS scheme it was demonstrated that by using pairs of laser pulses of different colors (e.g., red and blue) it is possible to excite a dilute ensemble of molecules such that lasing and/or gain-swept superradiance is realized in a direction back toward the out-going laser pulses (Fig. 7.10). This approach is a conceptual step toward spectroscopic probing at a distance, also known as SOS [18].

Another simpler approach was developed on the basis of the backward-directed lasing in optically excited plain air (Fig. 7.11). This technique relies on the remote generation of a weakly ionized plasma channel through filamentation of ultraintense trains of femtosecond laser pulses. Subsequent application of an energetic nanosecond pulse or series of pulses boosts the plasma density in the seed channel via avalanche ionization. Depending on the spectral and temporal content of the driving pulses, a transient population inversion is established in either nitrogen- or oxygen-ionized molecules, thus enabling a transient gain for an optical field propagating back toward the source and observer. This technique results in the generation of a strong, coherent, counter-propagating optical probe pulse. Such a probe, combined with a wavelength-tunable laser signal propagating in the forward direction, provides a tool for various remote sensing

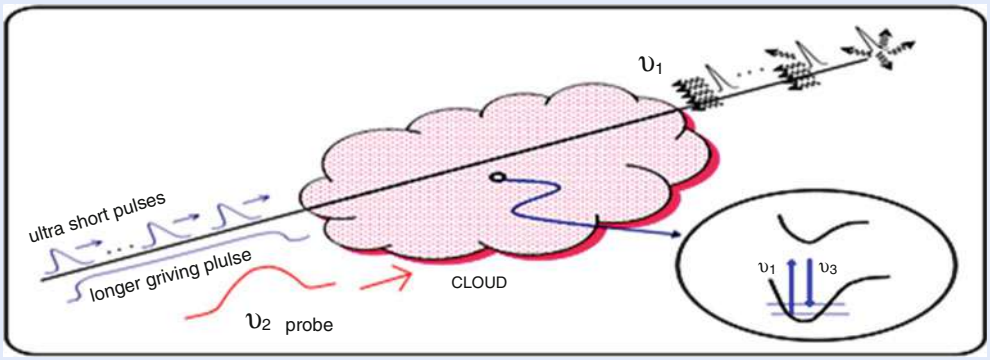


Fig. 7.11 SOS II. An ultrashort laser pulse or sequence of pulses is pre-chirped in such a way that they become compressed by the air dispersion at a pre-arranged distance behind the "cloud." Self-focusing collapse of these pulses results in the generation of weekly ionized "seed" plasma channels. The plasma density in these seed filaments is increased by several orders of magnitude by the application of a longer drive pulse. The properties of the pulses are tailored to produce population inversion in the ionized N_2 and O_2. The air laser radiation at frequency v_1 is combined with an interrogation pulse at v_2 to identify trace amounts of gas in the cloud. Adapted from [15]

applications. The technique could be implemented to probe the air directly above the growing crops and has the ability to pinpoint local areas of infection.

7.3.3 Detection of Plant Stress Using Laser-Induced Breakdown Spectroscopy

Remote sensing applied to detection of stress in plants is a very promising field. Plant stress affects the yield of agriculturally and economically significant plants. Rapid detection of plant stress in the field may allow farmers and crop growers to counter the effects of plant stress and increase their crop return. Plants can be affected by many types of stress including drought, pollution, insects, and microbial infestations which have a negative impact on agriculturally and economically significant florae. Rapid detection with little sample preparation is necessary to scale the sensing technology to the field size. We have recently applied laser-induced breakdown spectroscopy (LIBS) to plant stress detection. LIBS provides the advantages of rapid remote sensing [19]. LIBS measurements are performed on plants by focusing a laser pulse onto the surface of a sample causing ablation and vaporization of the sample material forming a plasma plume. The hot plasma with initial temperatures of up to 100,000 K expands and cools, emitting photons of characteristic frequencies from thermalized atomic constituents. LIBS can also be applied to detect other types of plant stress and other elements for phenotyping. For instance, cotton, a staple crop of Texas, is subject to various pestilences such as Southwestern rust (*Puccinia cacabata*), Alternaria leaf spot (*Alternaria macrospora*), and various types of boll rots. LIBS experiments can be performed using lab-based and portable femtosecond laser systems. We performed LIBS measurements for rapid analysis of the effects of drought stress on gardenia and wheat using a lab-based amplified femtosecond laser system operating at ~800 nm center wavelength, with a pulse duration of ~35 fs. We observed significant differences in the LIBS signals from stressed (not watered) and non-stressed (watered) plants and identified several atomic emission peaks as spectroscopic signatures of plant stress which agreed closely with macro- and micronutrients acquired by plants from the soil and air (Fig. 7.12). The LIBS technology may be able to identify key abiotic stress signals and improve the prediction of abiotic stress response capacity and can be used as a rapid remote sensing platform in the field.

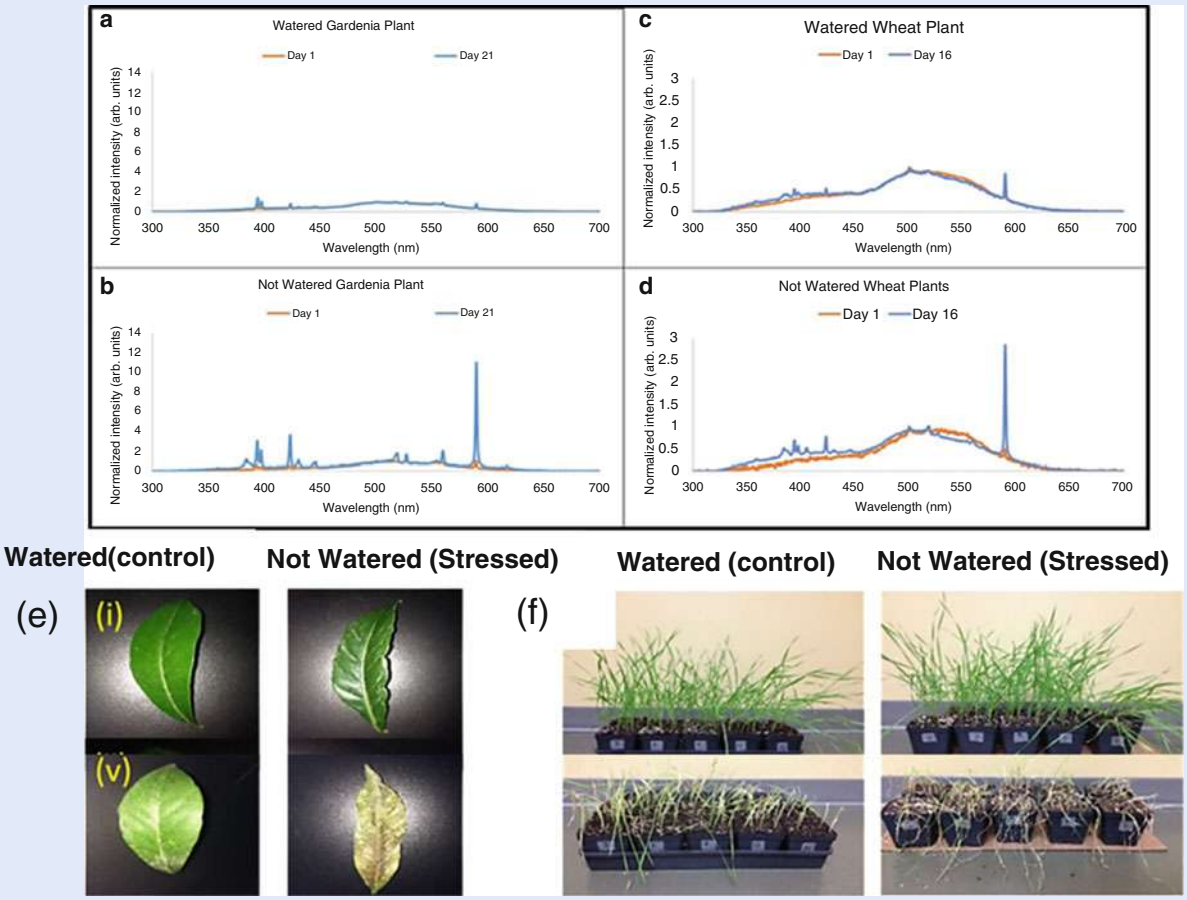


Fig. 7.12 (**a**–**d**) Averaged LIBS spectra of watered (non-stressed) and not watered (stressed) gardenia and wheat plants. The corresponding photographs of the gardenia leaves (**e**) and wheat plants (**f**) taken on the first and last day of the treatment

7.3.4 Stand-off Detection Using Laser Filaments

Femtosecond filamentation has been observed for various pulse durations (from several tens of femtoseconds to picoseconds) and wavelengths (from UV to IR). Due to the presence of high intensity electromagnetic field $I(r, t)$ the refractive index of the medium has the form $n = n_0 + n_2 I(r, t)$, where the nonlinear Kerr index n_2 leads to an interesting effect of the curvature of wave front acting like a focusing lens ($n_2 > 0$) since the intensity is usually the highest at the center of the beam. The latter leads to self-focusing which can overcome the diffraction and leads to the collapse if the input peak power P_{in} exceeds a critical threshold value P_{cr}. Filaments may be also used to induce rain via water condensation in the atmosphere and to induce lightning via electron condensation.

Filamentation of ultrashort laser pulses in the atmosphere offers unique opportunities for long-range transmission of high-power laser radiation and stand-off detection. With the critical power of self-focusing scaling as the laser wavelength squared, the quest for longer-wavelength drivers, which would radically increase the peak power and, hence, the laser energy in a single filament, has been ongoing over two decades, during which time the available laser sources limited filamentation experiments in the atmosphere to the near-infrared and

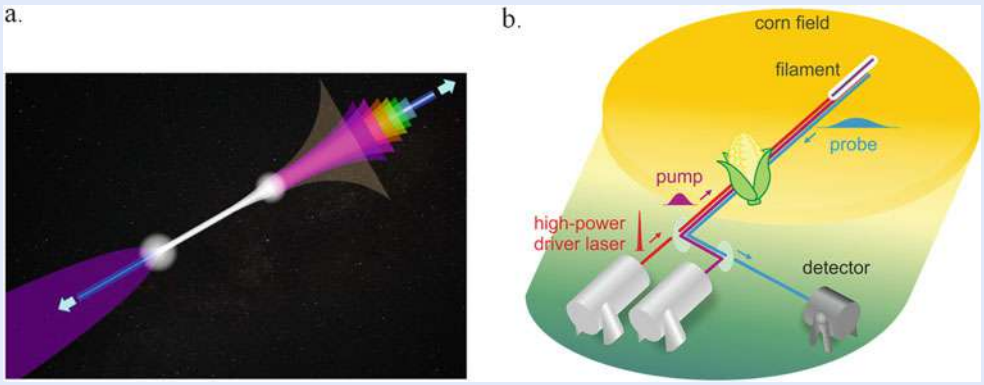


Fig. 7.13 (**a**) Laboratory prototype of a unique mobile high-power laser source of ultrashort pulses in the mid-infrared allows filamentation of ultrashort mid-infrared pulses in the atmosphere. (**b**) This innovative technology offers unprecedented opportunities for long-range signal transmission, delivery of high-power laser beams, and remote sensing in agricultural applications. Adapted from [20]

visible ranges. Recently, a unique high-power laser source of ultrashort pulses in the mid-infrared has been created, allowing filamentation of ultrashort mid-infrared pulses in the atmosphere to be demonstrated for the first time with the spectrum of a femtosecond laser driver centered at 3.9 μm, right at the edge of the atmospheric transmission window, radiation energies above 20 mJ and peak powers in excess of 200 GW that can be transmitted through the atmosphere in a single filament (Fig. 7.13) [20]. These studies revealed unique properties of mid-infrared filaments, where the generation of powerful mid-infrared supercontinuum is accompanied by unusual scenarios of optical harmonic generation, giving rise to remarkably broad radiation spectra, stretching from the visible to the mid-infrared. These thrilling discoveries open new horizons in ultrafast optical physics and offer unique opportunities for long-range signal transmission, delivery of high-power laser beams, and remote sensing of the atmosphere [21, 22]. Filament-induced breakdown spectroscopy may be realized as a remote extension of LIBS which was described in the previous section.

7.4 Quantum Heat Engines

Photonic quantum heat engines typically produce useful work by extracting energy from a high temperature thermal photon source, e.g., the Sun, and rejecting entropy to a low temperature entropy sink, e.g., the ambient room temperature surroundings. Lasing without inversion (LWI) is based on induced quantum coherence in the atoms, molecules, and solid state electrons involved. The early careful analyses of LWI were carried out by Kocharovskaya [23] using dark state, i.e., initial coherence; and Harris [24] using Fano interference, i.e., noise-induced coherence. Another related phenomenon is the correlated emission laser [25] which uses radiatively induced coherence to suppress absorption. It has also been shown that quantum coherence can be used to suppress absorption and obtain LWI [26].

More recently, it has become apparent that quantum coherence can be used to break detailed balance in a photocell and thus suppress recombination. This can increase quantum efficiency [27] and enhance thermodynamic power [28]. This reveals the deep connection between lasers and photovoltaic cells. It was shown that it is possible, in principle, to double the power of a thin medium photocell and/or the power of a laser. Finally it was also shown how quantum noise can be

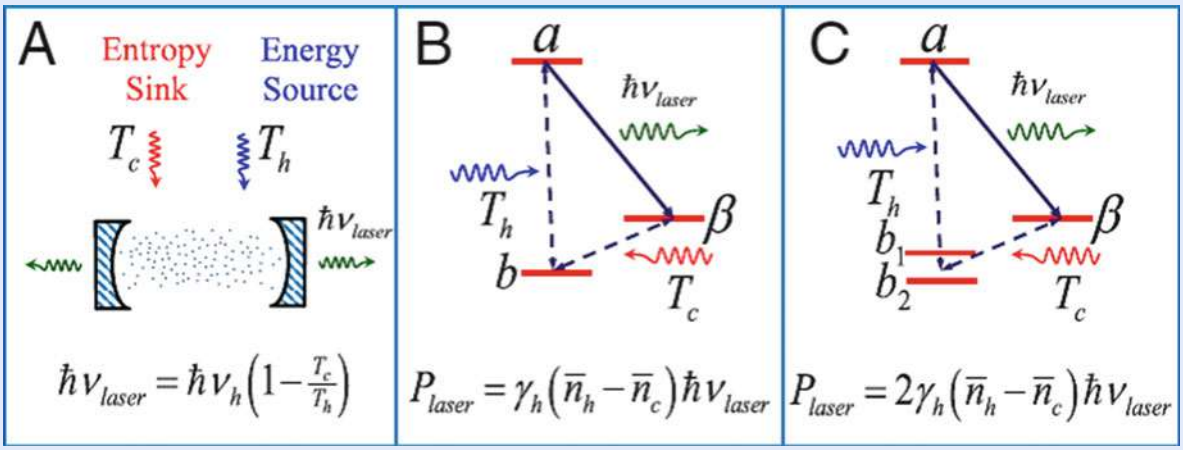


Fig. 7.14 (**a**) Schematic of a laser pumped by hot photons at temperature T_h (energy source, *blue*) and by cold photons at temperature T_c (entropy sink, *red*). The laser emits photons (*green*) such that at threshold the laser photon energy and pump photon energy is related by the Carnot efficiency relation. (**b**) Schematic of atoms inside the cavity. Lower level *b* is coupled to the excited states *a* and β. The laser power is governed by the average number of hot and cold thermal photons, $\bar{n}_h$ and $\bar{n}_c$. (**c**) Same as (**b**) but lower level *b* is replaced by two states b_1 and b_2, which can double the power when there is coherence between the levels. Adapted from [28]

used in the spirit of a quantum heat engine to explain the appearance of coherent oscillations and to increase of efficiency in photosynthesis [29].

7.4.1 The Laser and the Photovoltaic Cell as a Quantum Heat Engine

The arch-type example of a quantum heat engine is a laser pumped by hot thermal light as in Fig. 7.14. Then the frequency of a laser pumped by such narrow band hot light and cooled by narrow band cold light as well as the open circuit voltage of a solar cell or photo-detector obeys the Carnot relations.

However, it is possible to use quantum coherence to go beyond both of these Carnot relations. For example, including quantum coherence in the lower laser state, as in Fig. 7.14, yields the increase in quantum efficiency for the laser. Likewise it is possible to use quantum coherence to increase voltage quantum efficiency for the photovoltaic cell (Fig. 7.15). However one might well wonder if would that not violate the second law of thermodynamics? The answer is no. Quantum mechanics does allow us to get more energy from a thermal reservoir than a classical Carnot engine can, but at a cost. Overall the Carnot limit applies, but in a subsystem we can do better than the Carnot limit. A clean example of this is the photo-Carnot quantum heat engine which we discuss next.

7.4.2 The Photo-Carnot Quantum Heat Engine

The photo-Carnot engine is simply a piston engine in which photons replace the molecules as the driving fluid, as in Fig. 7.16. As such, the photons are like the steam molecules of a steam engine. However, the thermal photons are generated in

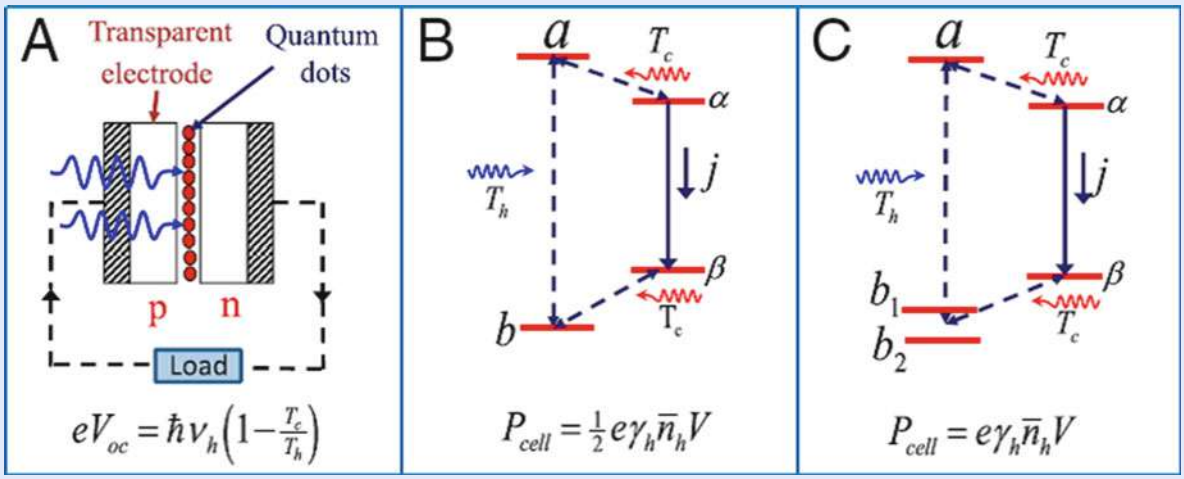


Fig. 7.15 (**a**) Schematic of a photocell consisting of quantum dots sandwiched between *p* and *n* doped semiconductors. Open circuit voltage and solar photon energy $\hbar\nu_h$ are related by the Carnot efficiency factor in which T_c is the ambient and T_h is the solar temperature. (**b**) Schematic of a *quantum dot* solar cell in which state *b* is coupled to *a* via, e.g., solar radiation and coupled to the conduction band reservoir state α via optical phonons. The electrons in state α pass to valence band reservoir state β via an external circuit, which contains the load. (**c**) Same as (**b**) but lower level *b* is replaced by two states b_1 and b_2, and when coherently prepared can double the output power. Adapted from [28]

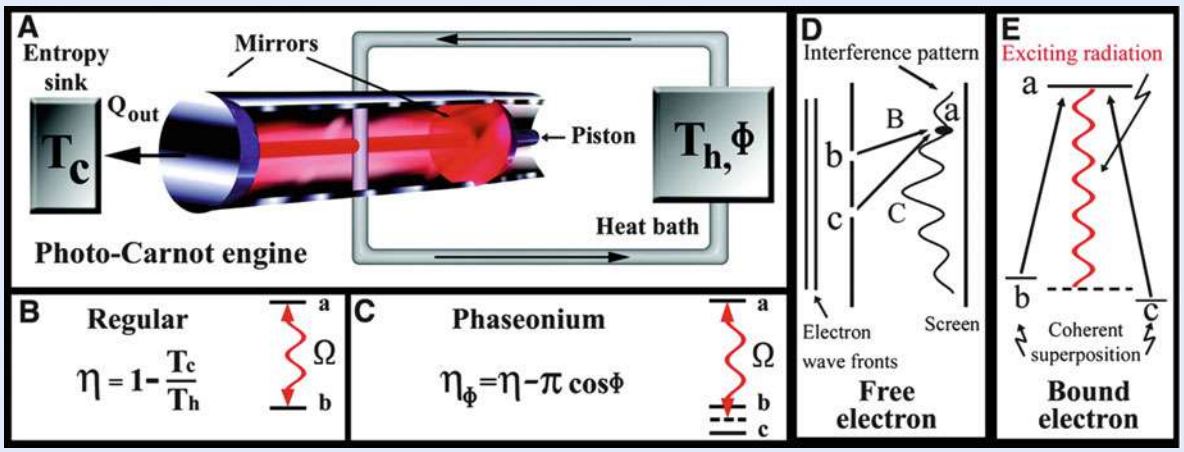


Fig. 7.16 (**a**) Photo-Carnot engine in which radiation pressure from a thermally excited single-mode field drives a piston. Atoms flow through the engine and keep the field at a constant temperature T_{rad} for the isothermal $1 \rightarrow 2$ portion of the Carnot cycle. Upon exiting the engine, the bath atoms are cooler than when they entered and are reheated by interactions with the hohlraum at T_h and "stored" in preparation for the next cycle. The combination of reheating and storing is depicted in (**a**) as the heat reservoir. A cold reservoir at T_c provides the entropy sink. (**b**) Two-level atoms in a regular thermal distribution, determined by temperature T_h, heat the driving radiation to $T_{rad} = T_h$ such that the regular operating efficiency is given by η. (**c**) When the field is heated, however, by a phaseonium in which the ground state doublet has a small amount of coherence and the populations of levels *a*, *b*, and *c* are thermally distributed, the field temperature is $T_{rad} > T_h$, and the operating efficiency is given by $\eta_\phi = \eta - \pi\cos(\phi)$. (**d**) A free electron propagates coherently from holes *b* and *c* with amplitudes *B* and *C* to point *a* on screen. The probability of the electron landing at point *a* shows the characteristic pattern of interference between (partially) coherent waves. (**e**) A bound atomic electron is excited by the radiation field from a coherent superposition of levels *b* and *c* with amplitudes *B* and *C* to level *a*. The probability of exciting the electron to level *a* displays the same kind of interference behavior as in the case of free electrons; i.e., as we change the relative phase between levels *b* and *c*, by, for example, changing the phase of the microwave field which prepares the coherence, the probability of exciting the atom varies sinusoidally. Adapted from [30]

the piston cylinder (electromagnetic cavity) of Fig. 7.16, by the hot atoms; that is, the atoms are the fuel, like the coal in a steam engine. Hence the classical photon driven heat engine must operate (at best) with Carnot thermodynamic efficiency since it is, at bottom, just another heat engine.

However, the plot thickens when we bring quantum coherence into the system. Now the (fuel) atoms can add more heat to the photon flux (working fluid) than was allowed by classical statistical mechanics, i.e., than was allowed by detailed balance. When we use quantum fuel (phaseonium) we can now "beat" the Carnot limit because we can now break detailed balance as per Fig. 7.16 [30]. This is a useful example to bear in mind when thinking about the quantum PV cell and photosynthesis.

7.4.3 Biological Quantum Heat Engines

Quantum entanglement [31] and other quantum coherence effects, e.g., the photon echo [32–36], have been investigated in a series of interesting photosynthesis experiments. The field of quantum biology of photosynthesis has been rapidly growing since the discovery of quantum coherence effects in the energy transfer process of the photosynthetic green sulfur bacteria [32, 35] and marine algae [33]. However, neither the role of quantum coherence nor the precise mechanism of the highly efficient energy transfer has been identified. Their description requires re-evaluation of the currently used methods and approximations. Also, whether the coherence is generated by coherent laser pulses used in the experiments or whether there is a kind of spontaneous coherence between the quantum levels involved as in the sense of noise-induced Fano-Agarwal interference has been the subject of debate. This latter possibility has been observed in various contexts and is indeed well known in quantum optics as described with applications to laser and solar cell quantum heat engines [27, 28, 37]. However it is not clear whether the quantum optical lore has applicability to photosynthesis for several reasons. High on the list being the question of environmental decoherence.

Recently, we have applied the formalism of the quantum heat engines to photosynthetic complexes such as light-harvesting antennae and reaction centers (RC) (Fig. 7.17). These systems operate as quantum heat engines and their structure is suitable to provide an increase in the efficiency of these processes. Connections between various quantum mechanical effects, namely the coherence/population coupling in photosynthesis [29, 38], and the quantum yield enhancement in laser and solar cell quantum heat engines [27, 28] were investigated. Analogy was drawn with a solar cell operation where the electron transfer efficiency may be increased by a quantum coherence between a doublet of closely lying states. It was proposed that the special pair of molecules in RC has a suitable structure to exhibit similar quantum effects. Figure 7.17 depicts the proposed schemes in which the efficiency of the electron transfer from the donor molecule (D) to the acceptor (A) may be increased by the quantum coherence between two donor molecules D_1 and D_2 of the special pair. This can provide insight into the structure–function relations of natural molecular architecture and will inspire new nature-mimicking artificial designs.

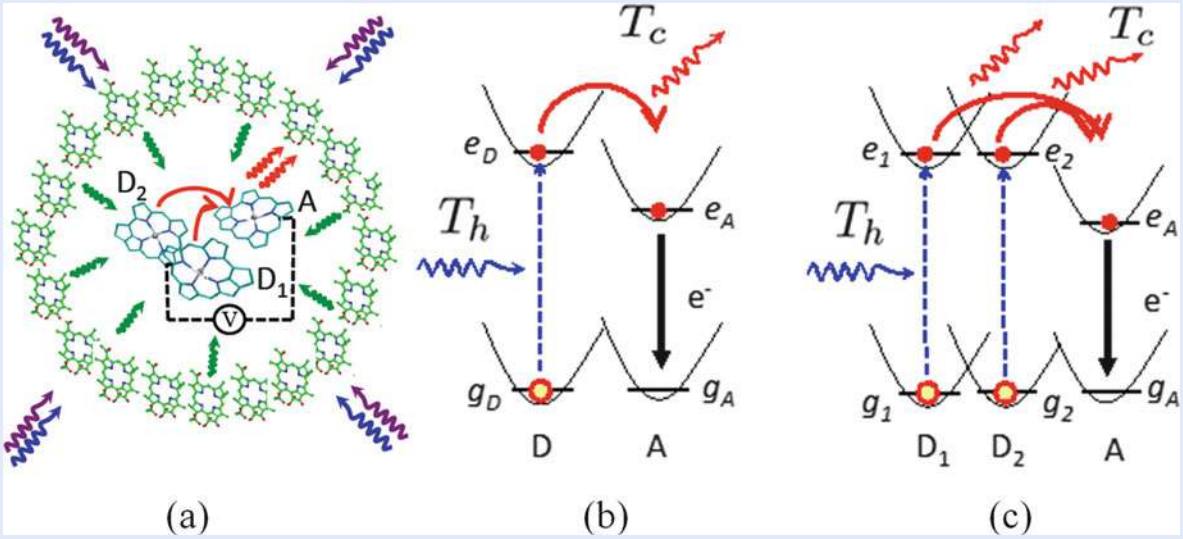


Fig. 7.17 Structure of a reaction center (RC) surrounded by photosynthetic antenna molecules (**a**). Schemes of the charge separation in the toy model of the biological quantum heat engine based on RC (**b**). The narrow band thermal radiation is transferred from the antennae complexes to the RC represented by donor (D) and acceptor (A) molecules. (**c**) Is the same as (**b**) with the upper level a is replaced by two levels a_1 and a_2. Quantum coherence between these levels can increase the power delivered by such a device. Adapted from [29]

7.5 Emerging Techniques with Single Molecule Sensitivity

7.5.1 Coherent Surface-Enhanced Raman Spectroscopy

Biological samples may be analyzed using fluorescence labeling nano-sensing technology such as enzyme-linked immunosorbent assay which is based on antibodies for detection of the presence of specific substances. The principle of this technique is based on the modification of the refractive index of the sensor when the antigen of interest is present. The sensitivity can be further enhanced by placing plasmonic nanoparticles so as to allow surface-enhanced Raman spectroscopy (SERS). These very sensitive techniques can be used to detect extremely small amounts of antigens.

Other techniques such as label-free optofluidic nanoplasmonic sensors have recently shown promise for detection of live viruses in biological media (Fig. 7.18) [39]. Such sensors are based on antibodies immobilized on plasmonic nanostructures such as arrays of small holes in gold or silver chips. These sensors show significantly modified properties of scattered or transmitted light due to the capture of viruses by the attached antibodies. This technology is promising for early diagnosis of pathogens from human blood.

Recently we developed a time-resolved surface-enhanced CARS (tr-SECARS) technique where the gain in signal was attributed to the enhanced electromagnetic fields that are created near the metal particles and in the gaps between the particles or features on the tailored substrates (Fig. 7.19) [8]. Thus the molecules of interest experience these enhanced fields by being attached to or simply near these metal particles or features. This coherent extension of the SERS technique provides an additional signal enhancement due to laser-induced vibrational coherence. It also provides the possibility to study biological systems simultaneously with a high

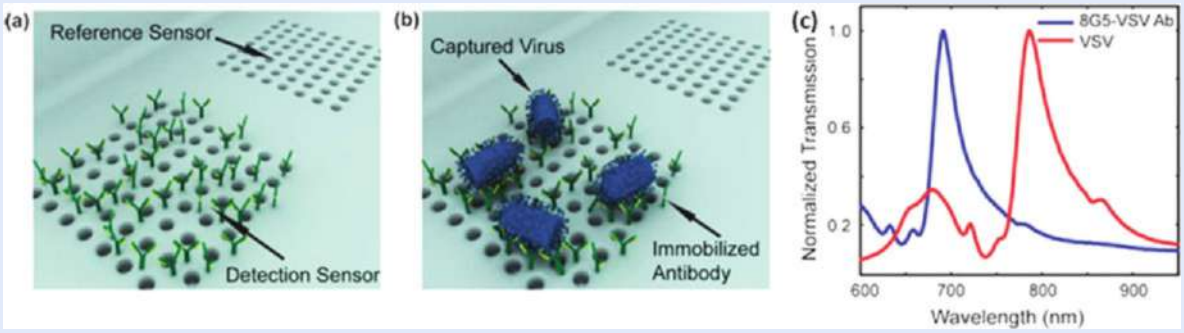


Fig. 7.18 Detection of viruses captured by immobilized antibodies on optofluidic nanoplasmonic sensors (**a**, **b**). Spectra of light transmitted by small holes change due to the capture of viruses (**c**). Adapted from [39]

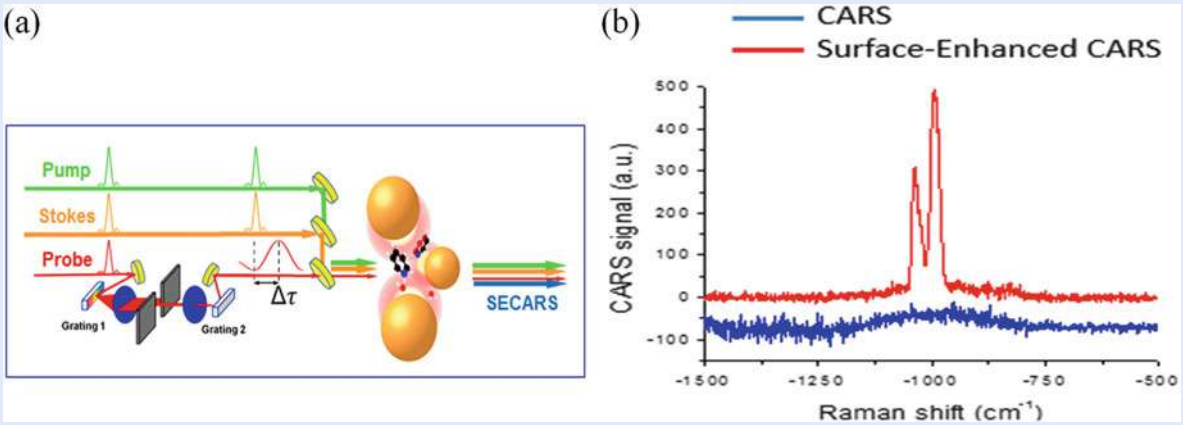
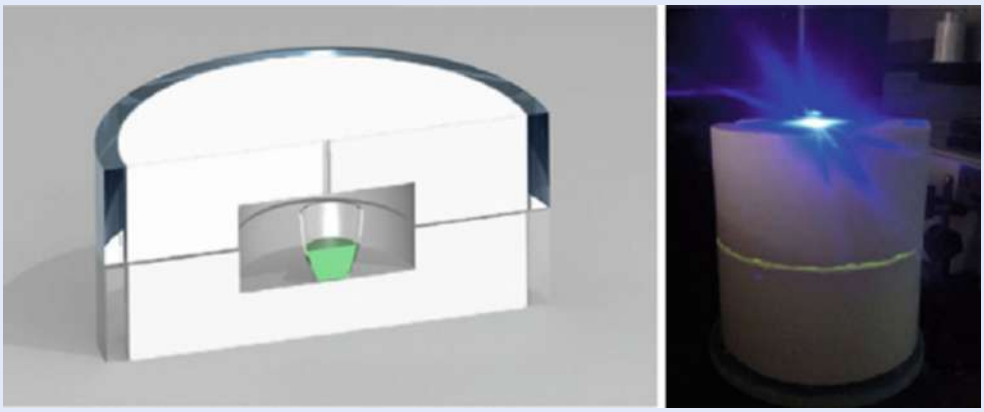


Fig. 7.19 (**a**) Experimental scheme of the time-resolved surface-enhanced anti-Stokes Raman scattering (tr-SECARS) spectroscopy. (**b**) Surface-enhanced CARS (*red*) reveals traces of hydrated pyridine molecules on the surface of gold nanoparticle aggregates with higher sensitivity than the conventional CARS (*blue*). Adapted from [8]

spatial and temporal resolution. Such a SECARS technique has the best-of-both-worlds combination of signal enhancements, i.e., the surface and coherence enhancements. This tr-SECARS technique increased the CARS signal intensity by seven orders of magnitude and was used to detect trace amounts of water on the surface of aggregated gold nanoparticles [8]. Conventional SERS was not able to detect any signal under the same conditions. This was recently reviewed in the literature as having "astonishing" sensitivity [9].

7.5.2 Cavity Ring-Down Spectroscopy

Food and water are essential for life, thus maintaining the safety of these resources is a high priority. The existing problem will only get worse in the years to come, due to limited fresh water supplies and increasing impacts of contaminants on the environment. Importantly, the ability to detect and manage environmental



7

Fig. 7.20 Integrating cavity setup for ultrasensitive detection of waste products such as urobilin in water. Adapted from [7]

contamination in food and water (in real time) is a critical need for the assessment and reduction of risk.

Traditional epi-illumination fluorescence spectroscopy systems use an objective lens to focus excitation light into the sample and collect the fluorescence emission. In such a configuration, the generated signal is limited to the focal volume of the optics, and is diffusive in nature; only a small fraction of the total emitted light is collected. Because only a small volume of a sample can be probed at any given time with such a configuration, detection of subnanomolar concentrations remains difficult. Thus, a method that could allow for exciting a larger volume of a sample while also providing means for collecting more of the fluorescence emission could greatly enhance the ability to detect sub-picomolar concentrations of, for example, urobilin (Fig. 7.20).

To achieve both of these goals, an integrating cavity was recently used to enhance both the excitation and collection efficiency [7]. Integrating cavities, especially spherical cavities, are commonly used to measure the total radiant flux from a source, as a means to generate uniform illumination, and as pump cavities for lasers. The high reflectivity of the cavity walls leads to a very large effective optical path length over which fluorescence excitation occurs; and the result is excitation of the entire volume of any sample placed inside the cavity. The high reflectivity of the cavity walls also means that fluorescence is collected from all directions; the result is the ability to detect a 500 femtomolar concentration of urobilin [7].

An accurate knowledge of the spectral absorption of cells and their constituents is critical to the continuation of progress in the understanding and modeling of biological and biomedical processes. Modeling of highly scattering cellular media, imaging of tissues, cells, and organelles, and laser-based surgical procedures are just a few of the vast number of techniques and procedures that rely on a refined understanding of biological absorption coefficients. Previously absorption measurements have generally been performed using transmission-style experiments in which an attenuation coefficient is measured by observing the decrease in the intensity of a light source as it passes through a sample. While the attenuation coefficient includes the losses due to absorption, it also includes the contribution from any scattering present in the sample. But these problems are avoided by using an integrating cavity for absorption measurements. Specifically, due to the nearly Lambertian behavior of the cavity walls, an isotropic field is created inside the cavity and scattering within the sample cannot change that. The result is absorption measurements that are independent of scattering. The integrating cavity technique was used to obtain what are now widely considered

to be the standard reference data for pure water absorption [40, 41]. Cavity ring-down spectroscopy (CRDS) is a technique developed for highly accurate and sensitive measurements of low absorption coefficients. It involves sending a temporally short laser pulse into a high-finesse two mirror cavity and observing the exponential decay, or "ring-down" of the intensity due to absorption (and scattering) inside the cavity. While CRDS is a very powerful technique, it is inherently unable to distinguish the losses due to absorption from losses due to scattering. Much like other transmission-style experiments, it is the attenuation that is being measured. This is clearly a problem for determining weak absorption in highly scattering samples. The new approach is based on CRDS and is called integrating cavity ring-down spectroscopy (ICRDS) where the traditional two mirror cavity used in CRDS is replaced with an integrating cavity. Again, the integrating cavity walls are fabricated from a highly diffuse reflecting material which creates an isotropic light field inside the cavity; the result is that a measurement of the decay time for the optical pulse is inherently insensitive to any scattering in the sample. Thus, ICRDS can provide a direct measurement of the absorption coefficient, even in the presence of strong scattering. However, ICRDS has not previously been implemented because a material with a sufficiently high diffuse reflectivity was not available. Spectralon has been the material with the highest known diffuse reflectivity (99.3 %) [42], but at this level of reflectivity, the decay of the optical pulse due to losses during reflection from the wall is so large that the sensitivity to absorption is low.

A new material has now been developed that has a diffuse reflectivity up to 99.92 % [43]; this is sufficient to make ICRDS a reality. It opens up a plethora of exciting new applications with tremendous impact in, for example, the biomedical area where highly accurate absorption spectra of living cells, tissue samples, liquids, etc., can now be obtained, even when the absorption is very weak compared to scattering in the sample.

As a specific such example, consider human retinal pigmented epithelium (RPE) cells. Among other functions, RPE cells are responsible for absorbing scattered light to improve the optical system and reduce stress on the retina. Figure 7.21 shows an example in which ICRDS was used to measure the base absorption (i.e., with the pigment removed) of 60 million RPE cells in a 3 mL solution [44]. Also shown is the attenuation (scattering plus absorption) spectrum obtained with the same sample in a spectrophotometer. As an example of the dominance of scattering in the transmission measurement, consider the

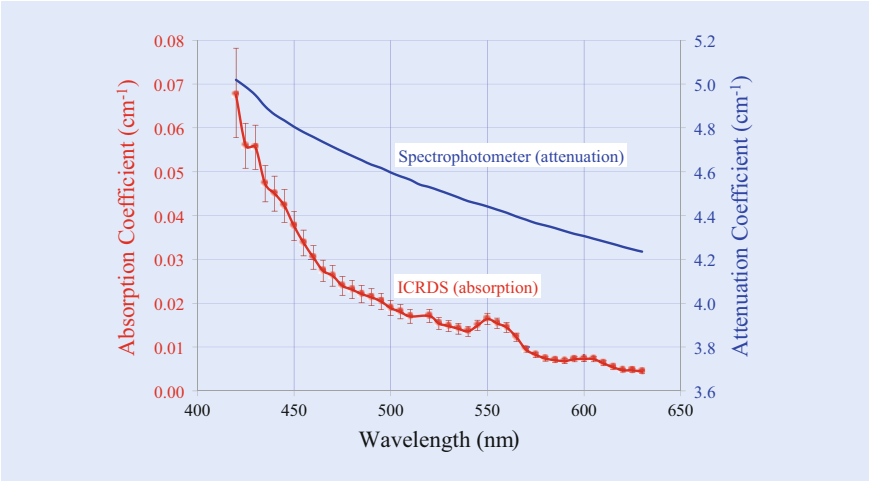


Fig. 7.21 Absorption coefficient of 6×10^7 human retinal pigmented epithelium (RPE) cells in a 3 mL solution; also shown is the attenuation (scattering + absorption) spectrum. Adapted from [44]

attenuation at 500 nm (4.6 cm^{-1}) and the absorption (~0.02 cm^{-1}). This is a factor of 230; the absorption is less than one-half of 1 % of the attenuation. Figure 7.21 shows biological absorption spectra that were previously inaccessible for study and analysis.

7.6 Superresolution Quantum Microscopy

7.6.1 Subwavelength Quantum Microscopy

The measurement of small distances is a fundamental problem of interest since the early days of science. It has become even more important due to recent interest in nanoscopic and mesoscopic phenomena in biophysics. Starting from the invention of the optical microscope around 400 years ago, today's optical microscopy methodologies can basically be divided into lens-based and lensless imaging. In general, far-field imaging is lens-based and thus limited by criteria such as the Rayleigh diffraction limit which states that the achievable resolution in the focus plane is limited to approximately half of the wavelength of illuminating light. Further limitation arises from out-of-focus light, which affects the resolution in the direction perpendicular to the focal plane.

Many methods have been suggested to break these limits. Lens-based techniques include confocal, nonlinear femtosecond, or stimulated emission depletion microscopy which have been recognized by the Nobel Prize. They have achieved remarkable first results, as shown in Fig. 7.1. Also non-classical features such as entanglement, quantum interferometry, or multi-photon processes can be used to enhance resolution. However, there is still great interest in achieving nanometer distance measurements by using optical illuminating far-field imaging only.

Recently, new schemes were proposed [45–47] to measure the distance between two adjacent two-level systems by driving them with a standing wave laser field and measuring the far-field resonance fluorescence spectrum, which is motivated by the localization of single atom inside a standing wave field to distances smaller than the Rayleigh limit $\lambda/2$. The basic idea is that in a standing wave, the effective driving field strength depends on the position of the particles (Fig. 7.22). Thus, each particle generates a sharp sideband peak in the spectrum, where the peak position directly relates to the subwavelength position of the particle. As long as the two sideband peaks can be distinguished from each other, the position of each particle can be recovered. However, when the interatomic distance decreases, the two particles can no longer be considered independent. Due to the increasing dipole–dipole interaction between the two particles, the fluorescence spectrum becomes complicated. It was found, however, that the

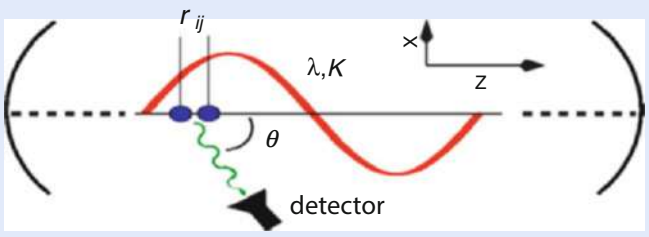


Fig. 7.22 Two atoms in a standing wave field separated by a distance r_{ij} smaller than half of the wavelength λ of the driving field. The distance of the two atoms is measured via the emitted resonance fluorescence

dipole–dipole interaction energy can directly be extracted from the fluorescence spectrum by adjusting the parameters of the driving field. Since the dipole–dipole interaction energy is distance dependent, it yields the desired distance information. The new schemes showed the applicability to inter-particle distances in a very wide range from $\lambda/2$ to about $\lambda/550$. These schemes can be extended for sensing the inter-molecular distances with applications in molecular and cellular biology, microbiology, and medical research leading to precision resolution in microorganisms such as protozoa, bacteria, molds, and sperms. Other applications would involve mapping of DNA.

7.6.2 Tip-Enhanced Quantum Bioimaging

Understanding the structure–function relationships of molecular constituents of biological pathogens is important for building more precise models and for the design of bacteria inactivating drugs, etc. Pathogenic diseases that are caused by viral pathogens include smallpox, influenza, mumps, measles, chickenpox, ebola, etc. New technology is needed for rapid detection and treatment, and for detailed studies of nanoscale components of pathogens with high spatial and temporal resolution and with simultaneous chemical analysis. Various nano-spectroscopic methods have been used to model pathogens. Direct imaging of the structural dynamics is challenging due to the small size of the biomolecules. Novel quantum optical techniques based on nanoscale sensors with high spatial and molecular-level resolution aim to improve the existing pathogen structure–property models.

Recent exceptional breakthroughs in the spatial resolution (<1 nm) make tip-enhanced Raman scattering (TERS), which is an important variation of SERS, a very powerful tool for in situ chemical analysis on the nanoscale. TERS has recently been applied to imaging biological systems. However, several challenges remain due to weak Raman signals and motion of live cells during long acquisition time imaging. We have performed such nanoantenna-tip-induced bio-sensing of bacteria and nanoscale imaging of biological cells (Fig. 7.23). The goal is to map the molecule–substrate interactions with nanoscale molecular-level spatial resolution in both topography (using atomic force microscopy (AFM)) and chemical identification (using Raman spectroscopy). Both electromagnetic and chemical enhancement effects can be used to identify the molecular biomarkers and nanoscale chemical surface properties of biological systems.

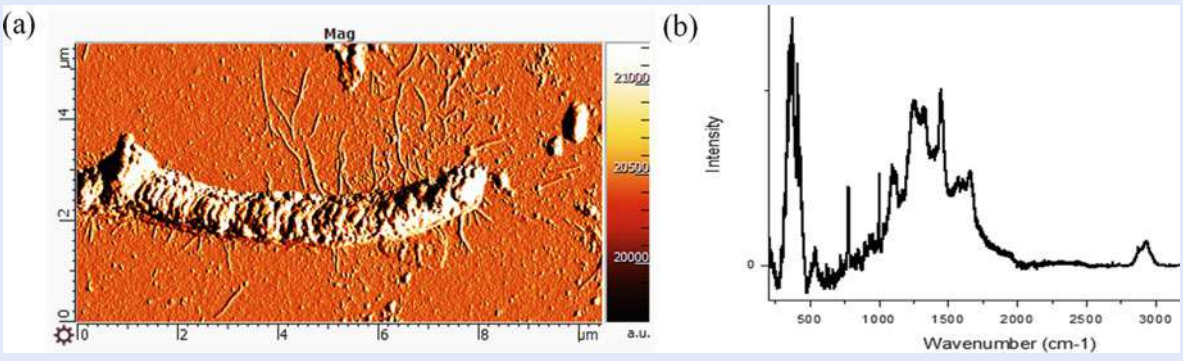


Fig. 7.23 (**a**) AFM image of a bacterial cell. (**b**) Typical Raman spectrum of bacteria

7.7 Novel Light Sources

7.7.1 Fiber Sensors

Recent research on fiber sensing has been focused on the development and demonstration of advanced fiber components and fiber-based strategies for quantum sensing. Fiber-bundle microprobe sensors have been developed using specifically designed fiber bundles and coupled to a confocal optical microscope to enable multiplex sensing, including multicolor in vivo detection of neuronal activity in a living brain using fluorescent protein biomarkers. Fiber-bundle microprobe sensors have also been used to confront the long-standing issues in practical imaging and sensing, including fiber-based endoscopic Raman imaging (Fig. 7.24).

Recent advances in optical magnetometry pave the ways toward an unprecedented spatial resolution and a remarkably high sensitivity in magnetic field detection, offering unique tools for the measurement of weak magnetic fields in a broad variety of areas from astrophysics, geosciences, and the physics of fundamental symmetries to medicine and life sciences. To unleash the full potential of this emerging technology and make it compatible with the requirements of practical quantum technologies and in vivo studies in life sciences, optical magnetometers have to be integrated with fiber-optic probes. This challenge has been addressed by developing a scanned fiber-optic probe for magnetic field imaging where nitrogen–vacancy (NV) centers are coupled to an optical fiber integrated with a two-wire microwave transmission line. The electron spin of NV centers in a diamond microcrystal attached to the tip of the fiber probe is manipulated by a frequency-modulated microwave field and is initialized by laser radiation transmitted through the optical tract of the fiber probe (Fig. 7.25) [50, 51]. The photoluminescence spin-readout return from NV centers is captured and delivered by the same optical fiber, allowing the two-dimensional profile of the magnetic field to be imaged with high speed and high sensitivity.

7.7.2 Quantum Coherence in X-Ray Laser Generation

The application of the techniques of quantum coherence and LWI to areas such as XUV and X-ray laser generation holds promise. The quantum coherence in atomic

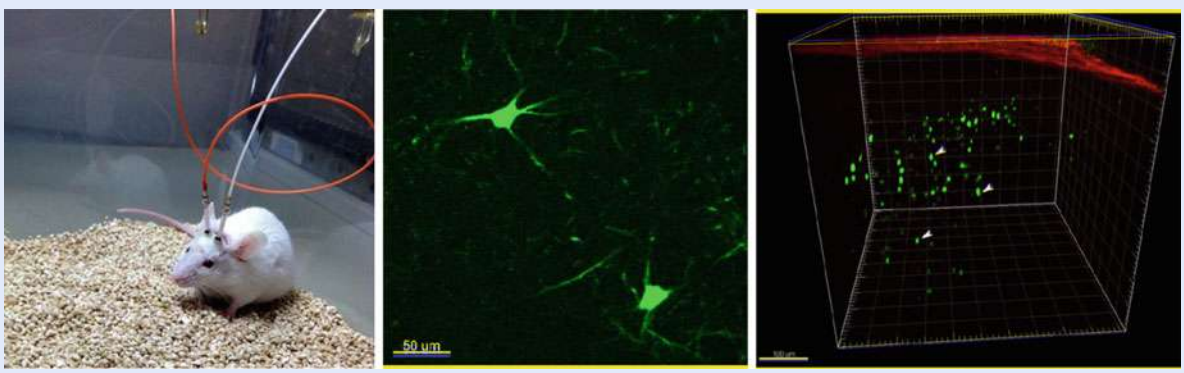


Fig. 7.24 Fiber-based sensing and imaging for neurophotonics and agricultural applications. Adapted from [48, 49]

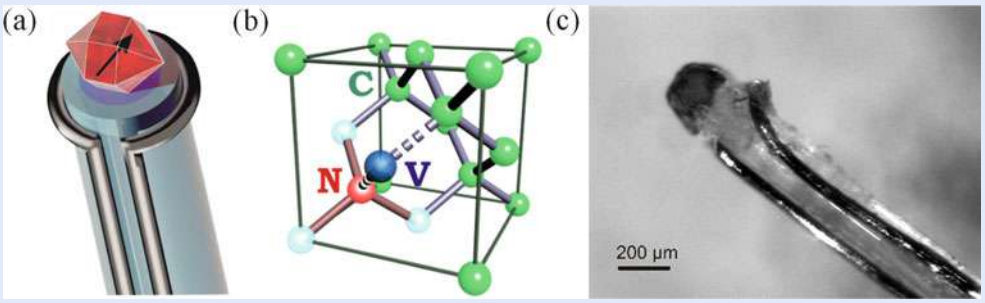


Fig. 7.25 (**a**) A fiber probe integrating an NV-diamond quantum sensor, an optical fiber, and a microwave transmission line. (**b**) A nitrogen atom (N) and a vacancy (V) forming an NV center in a diamond lattice, consisting of carbon (C) atoms. (**c**) Image of a prototype fiber-optic magnetometer and thermometer. Adapted from [50, 51]

and radiation physics has led to many interesting and unexpected consequences. For example, an atomic ensemble prepared in a coherent superposition of states yields self-induced transparency, photon echo, and coherent Raman beats [52, 53]. Another important coherence effect in atoms leads to phenomena of electromagnetically induced transparency, where by preparing an atomic system in a coherent superposition of states, under certain conditions, it is possible for atomic coherence to cancel absorption but not emission (Fig. 7.26). This is the basis for LWI, the essential idea being the absorption cancellation by atomic coherence and interference. Frequently this is accomplished in three- or four-level atomic systems in which there are coherent routes for absorption that can destructively interfere, thus leading to the cancellation of absorption. A small population in the excited state can thus lead to net gain, and this was the subject of substantial theoretical work by us several groups in the 1980s [24–26]. We have carried out the first LWI demonstrations in the mid-1990s [54, 55], and have recently continued in theoretical and experimental fronts to investigate the possibility of coherence driven lasing and lasing in XUV and soft X-ray regions [56, 57].

Existing X-ray laser sources such as X-ray free electron laser (X-FEL) at SLAC provided tremendous excitement for scientists in various disciplines. However, very large cost and size of this great device motivates researchers to search for portable, inexpensive XUV (see for example Suckewer, et al. [56]), and X-ray devices. On the other hand, the table top soft X-ray (SXL) and XUV lasers, due to their compactness, excellent beam quality, and very reliable operation in wavelength range of 10–50 nm hold great promise for tools for high resolution microscopy, micro-holography, very high plasma density measurements, semiconductor surface studies, and nano-lithography. Intensive efforts have been made to develop such compact soft X-ray lasers that are suitable for applications in academic and industrial laboratories as well as in the field biophotonics applications. LWI in the X-ray region would provide appealing opportunities and profound impacts on X-ray laser science as well as studies in the fields such as crystallography, solid materials, health sciences, high resolution microscopy of biological systems, and many more.

7.7.3 Coherent Control of Gamma Rays

Active control of light–matter interactions is an ultimate goal of many quantum biophotonics applications. Coherent control by laser pulse shaping has been a highly active research area for the last two decades addressing a large variety of problems including control of chemical reactions, spectroscopic signals, optical

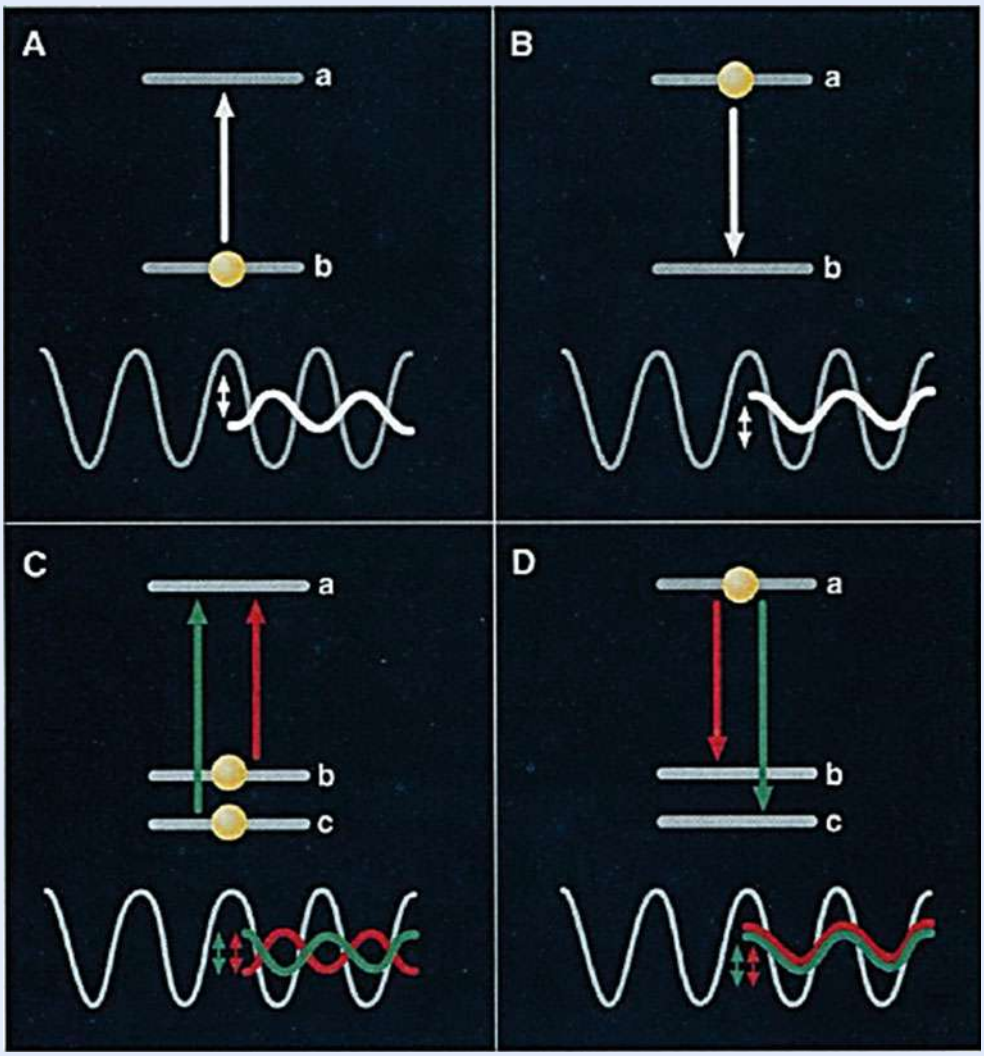


Fig. 7.26 Lasing without inversion (LWI) first experimentally realized at TAMU [54, 55]. Absorption cancellation by atomic coherence and interference is the basis for LWI. Adapted from [58]

sensors, microscopic images, photolithography, optical nanostructures, quantum logic operations, etc. All these applications, however, were aimed at controlling electronic motions, molecular vibrations and rotations using shaped electromagnetic radiation in the frequency range spanning from infrared to optical to ultraviolet range. The X-ray and gamma ray sources with controllable waveforms were up till now unavailable, and therefore there were no related applications.

Now, for the first time, we have an available source of controllable gamma rays [59]. Single ultrashort gamma ray pulses, double pulses, and pulse trains with controllable pulse delays can be produced (Fig. 7.27). Splitting of the single photon into two pulses constitutes the first realization of a time-bin qubit in this range of frequencies. The production of single-photon ultrashort-pulse trains with pulse duration much shorter than the natural lifetime of the emitting nuclear level and with the controllable waveforms provides a unique opportunity for the realization of quantum memories and other nuclear-ensemble—gamma-photons interfaces, opening the prospects for fascinating applications in quantum communication and information.

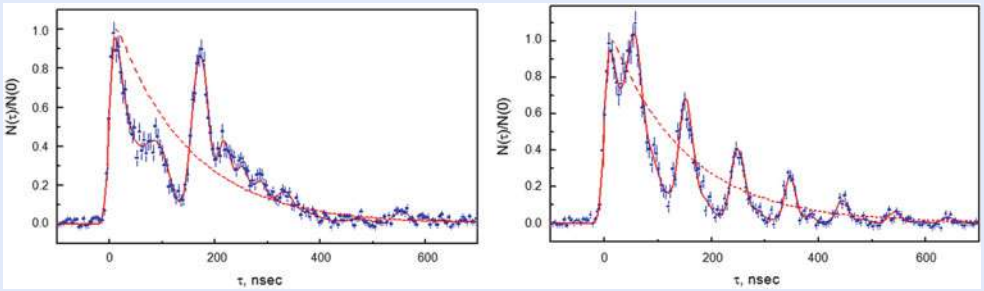


Fig. 7.27 Splitting gamma ray photons into double pulses (*left*) and multiple pulse trains (*right*). Adapted from [59]

The energy of these gamma ray pulses corresponds to nuclear transitions. This may enable, for the first time, coherent control of nuclear reactions. The parameters of the pulses can be widely controlled including the shape and number of pulses, repetition rate (potentially variable in the range from MHz to GHz), and duration (potentially ranged from 100 ns to 100 ps). This allows for a variety of time-resolved experiments (including dynamic X-ray diffraction). Gamma rays can provide an extremely high (potentially sub-Angstrom and currently nanometer) spatial resolution. The simultaneous high spatial and temporal resolution makes this technology promising for nanoscale ultrafast imaging of electronic motions in biomolecules and protein folding dynamics.

7.8 Conclusion

In summary, we have (hopefully) shown that at the interface between quantum optics and biophysics lies the emerging exciting field of quantum biophotonics. With new light sources and quantum techniques, it becomes increasingly possible to apply the techniques of quantum optics to biosciences. This will have major payoff in agriculture, environmental science, and national security, and promises to be the dawn of a new era in biophotonics emphasizing quantum effects.

Acknowledgment We acknowledge the support of the National Science Foundation Grants No. EEC-0540832 (MIRTHE ERC), No.PHY-1068554, No. PHY-1241032 (INSPIRE CREATIV), No. DBI-1455671, No. DBI-1532188, No. ECCS-1509268 and No. PHY-1307153, the Office of Naval Research grant N00014-16-1-3054, the US DOD awards FA9550-15-1-0517 and N00014-16-1-2578, and CPRIT grant RP160834, and the Robert A. Welch Foundation (Awards A-1261 and A-1547).

References

1. Wildanger D, Rittweger E, Kastrup L, Hell SW (2008) STED microscopy with a supercontinuum laser source. Opt Express 16:9614
2. Chmyrov A, Keller J, Grotjohann T, Ratz M, d'Este E, Jakobs S, Eggeling C, Hell SW (2013) Nanoscopy with more than 100,000 'doughnuts'. Nat Methods 10:737
3. Scully MO, Kattawar GW, Lucht RP, Opatrny T, Pilloff H, Rebane A, Sokolov AV, Zubairy MS (2002) FAST CARS: Engineering a laser spectroscopic technique for rapid identification of bacterial spores. Proc Natl Acad Sci U S A 99:10994
4. Pestov D, Murawski RK, Ariunbold GO, Wang X, Zhi M, Sokolov AV, Sautenkov VA, Rostovtsev YV, Dogariu A, Huang Y, Scully MO (2007) Optimizing the laser-pulse configuration for coherent Raman spectroscopy. Science 316:265
5. Pestov D, Wang X, Ariunbold GO, Murawski RK, Sautenkov VA, Dogariu A, Sokolov AV, Scully MO (2008) Single-shot detection of bacterial endospores via coherent Raman spectroscopy. Proc Natl Acad Sci U S A 105:422
6. Arora R, Petrov GI, Yakovlev VV, Scully MO (2012) Detecting anthrax in the mail by coherent Raman microspectroscopy. Proc Natl Acad Sci U S A 109:1151
7. Bixler LN, Cone MT, Hokr BH, Mason JD, Figueroa E, Fry ES, Yakovlev VV, Scully MO (2014) Ultrasensitive detection of waste products in water using fluorescence emission cavity-enhanced spectroscopy. Proc Natl Acad Sci U S A 111:7208
8. Voronine DV, Sinyukov AM, Xia H, Wang K, Jha PK, Welch G, Sokolov AV, Scully MO (2012) Time-resolved surface-enhanced coherent sensing of nanoscale molecular complexes. Sci Rep 2:891
9. Lis D, Cecchet F (2014) Localized surface plasmon resonances in nanostructures to enhance nonlinear vibrational spectroscopies: towards an astonishing molecular sensitivity. Beilstein J Nanotechnol 28:2275
10. Altangerel N, Ariunbold G, Gorman C, Bohlmeyer D, Yuan J, Hemmer P, Scully MO (2016) Early, in vivo, detection of abiotic plant stress responses via Raman spectroscopy. Conference on Lasers and Electro-Optics OSA Technical Digest (Optical Society of America, 2016), paper SF1H.3. doi:► 10.1364/CLEO_SI.2016.SF1H.3
11. Palombo F, Madami M, Stone N, Fioretto D (2014) Mechanical mapping with chemical specificity by confocal Brillouin and Raman microscopy. Analyst 139:729
12. Traverso AJ, Thompson JV, Steelman ZA, Meng Z, Scully MO, Yakovlev VV (2015) Dual Raman-Brillouin microscope for chemical and mechanical characterization and imaging. Anal Chem 87:7519
13. Meng Z, Traverso AJ, Yakovlev VV (2014) Background clean-up in Brillouin microspectroscopy of scattering medium. Opt Express 22:5410
14. Petrov GI, Arora R, Yakovlev VV, Wang X, Sokolov AV, Scully MO (2007) Comparison of coherent and spontaneous Raman microspectroscopies for noninvasive detection of single bacterial endospores. Proc Natl Acad Sci U S A 104:7776
15. Hemmer PR, Miles RB, Polynkin P, Siebert T, Sokolov AV, Sprangle P, Scully MO (2011) Standoff spectroscopy via remote generation of a backward-propagating laser beam. Proc Natl Acad Sci U S A 108:3130
16. Dogariu A, Michael JB, Scully MO, Miles RB (2011) High-gain backward lasing in air. Science 331:442
17. Traverso AJ, Sanchez-Gonzalez R, Yuan L, Wang K, Voronine DV, Zheltikov AM, Rostovtsev Y, Sautenkov VA, Sokolov AV, North SW, Scully MO (2012) Coherence brightened laser source for atmospheric remote sensing. Proc Natl Acad Sci U S A 109:15185
18. Kocharovsky V, Cameron S, Lehmann K, Lucht R, Miles R, Rostovtsev Y, Warren W, Welch GR, Scully MO (2005) Gain-swept superradiance applied to the stand-off detection of trace impurities in the atmosphere. Proc Natl Acad Sci U S A 102:7806
19. Cremers DA, Radziemski LJ (2006) Handbook of laser-induced breakdown spectroscopy 302. John Wiley, West Sussex
20. Mitrofanov AV, Voronin AA, Sidorov-Biryukov DA, Pugžlys A, Stepanov EA, Andriukaitis G, Flöry T, Ališauskas T, Fedotov AB, Baltuška A, Zheltikov AM (2014) Mid-infrared laser filaments in the atmosphere. Sci Rep 5:8368

21. Malevich PN, Kartashov D, Ališauskas ZPS, Pugžlys A, Baltuška A, Giniūnas L, Danielius R, Lanin AA, Zheltikov AM, Marangoni M, Cerullo G (2012) Ultrafast-laser-induced backward stimulated Raman scattering for tracing atmospheric gases. Opt Express 20:18784
22. Malevich PN, Maurer R, Kartashov D, Ališauskas S, Lanin AA, Zheltikov AM, Marangoni M, Cerullo G, Baltuška A, Pugžlys A (2015) Stimulated Raman gas sensing by backward UV lasing from a femtosecond filament. Opt Lett 40:2469
23. Kocharovskaya O, Khanin YI (1988) Coherent amplification of an ultrashort pulse in a three-level medium without a population inversion. JETP Lett 48:630
24. Harris SE (1989) Lasers without inversion: interference of lifetime-broadened resonances. Phys Rev Lett 62:1033
25. Scully MO (1985) Correlated spontaneous-emission lasers: quenching of quantum fluctuations in the relative phase angle. Phys Rev Lett 55:2802
26. Scully MO, Zhu SY, Gavrielides A (1989) Degenerate quantum-beat laser: lasing without inversion and inversion without lasing. Phys Rev Lett 62:2813
27. Scully MO (2010) Quantum photocell: using quantum coherence to reduce radiative recombination and increase efficiency. Phys Rev Lett 104:207701
28. Scully MO, Chapin KR, Dorfman KE, Kim M, Svidzinsky AA (2011) Quantum heat engine power can be increased by noise-induced coherence. PNAS 108:15097
29. Dorfman KE, Voronine DV, Mukamel S, Scully MO (2013) Photosynthetic reaction center as a quantum heat engine. PNAS 110:2746
30. Scully MO, Zubairy MS, Agarwal GS, Walther H (2003) Extracting work from a single heat bath via vanishing quantum coherence. Science 299:862
31. Sarovar M, Ishizaki A, Fleming GR, Whaley KB (2010) Quantum entanglement in photosynthetic light-harvesting complexes. Nat Phys 6:462
32. Engel GS, Calhoun TR, Read EL, Ahn TK, Mancal T, Cheng YC, Blankenship RE, Fleming GR (2007) Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. Nature 446:782
33. Collini E, Wong CY, Wilk KE, Curmi PM, Brumer P, Scholes GD (2010) Coherently wired light-harvesting in photosynthetic marine algae at ambient temperature. Nature 463:644
34. Panitchayangkoon G, Hayes D, Fransted KA, Caram JR, Harel E, Wen J, Blankenship RE, Engel GS (2010) Long-lived quantum coherence in photosynthetic complexes at physiological temperature. Proc Natl Acad Sci U S A 107:12766
35. Brixner T, Stenger J, Vaswani HM, Cho M, Blankenship RE, Fleming GR (2005) Two-dimensional spectroscopy of electronic couplings in photosynthesis. Nature 434:625
36. Abramavicius D, Palmieri B, Voronine DV, Sanda F, Mukamel S (2009) Coherent multidimensional optical spectroscopy of excitons in molecular aggregates; quasiparticle versus supermolecule perspectives. Chem Rev 109:2350
37. Kozlov VV, Rostovtsev Y, Scully MO (2006) Inducing quantum coherence via decays and incoherent pumping with application to population trapping, lasing without inversion, and quenching of spontaneous emission. Phys Rev A 74:063829
38. Panitchayangkoon G, Voronine DV, Abramavicius D, Mukamel D, Engel GS (2011) Direct evidence of quantum transport in photosynthetic light harvesting complexes. PNAS 108:20908
39. Yanik AA, Huang M, Kamohara O, Artar A, Geisbert TW, Connor JH, Altug H (2010) An optofluidic nanoplasmonic biosensor for direct detection of live viruses from biological media. Nano Lett 10:4962
40. Fry ES, Kattawar GW, Pope RM (1992) Integrating cavity absorption meter. Appl Opt 31:2055
41. Pope RM, Fry ES (1997) Absorption spectrum (380-700 nm) of pure water: II. Integrating cavity measurements. Appl Opt 36:8710
42. Labsphere (2006) A guide to reflectance coatings and materials. Tech. Rep. ► http://www.labsphere.com
43. Cone MT, Musser JA, Figueroa E, Mason JD, Fry ES (2015) Diffuse reflecting material for integrating cavity spectroscopy—including ring-down spectroscopy. Appl Opt 54:334
44. Cone MT, Mason JD, Figueroa E, Hokr BH, Bixler JN, Castellanos CC, Wigle JC, Noojin GD, Rockwell BA, Yakovlev VV, Fry ES (2015) Measuring the absorption coefficient of biological materials using integrating cavity ring-down spectroscopy. Optica 2:162
45. Chang JT, Evers J, Scully MO, Zubairy MS (2006) Measurement of the separation between atoms beyond diffraction limit. Phys Rev A 73:031803
46. Chang JT, Evers J, Zubairy MS (2006) Distilling two-atom distance information from intensity-intensity correlation functions. Phys Rev A 74:043820
47. Liao Z, Alamri M, Zubairy MS (2012) Resonance-fluorescence-localization microscopy with subwavelength resolution. Phys Rev A 85:023810

48. Doronina-Amitonova LV, Fedotov IV, Ivashkina OI, Zots MA, Fedotov AB, Anokhin KV, Zheltikov AM (2013) Implantable fiber-optic interface for parallel multisite long-term optical dynamic brain interrogation in freely moving mice. Sci Rep 3:3265
49. Doronina-Amitonova LV, Fedotov IV, Fedotov AB, Anokhin KV, Zheltikov AM (2015) Neurophotonics: optical methods to study and control the brain. Phys Usp 58:345
50. Fedotov IV, Safronov NA, Shandarov YA, Lanin AA, Fedotov AB, Kilin SY, Sakoda K, Scully MO, Zheltikov AM (2012) Guided-wave-coupled nitrogen vacancies in nanodiamond-doped photonic-crystal fibers. Appl Phys Lett 101:031106
51. Fedotov IV, Doronina-Amitonova LV, Voronin AA, Levchenko AO, Zibrov SA, Sidorov-Biryukov AD, Fedotov AB, Velichansky VL, Zheltikov AM (2014) Electron spin manipulation and readout through an optical fiber. Sci Rep 4:5362
52. Zhu SY, Nikonov DE, Scully MO (1998) A scheme for noninversion lasing for short-wavelength lasers in helium like ions. Found Phys 28:611
53. Rostovtsev Y, Scully MO (2007) Soft X-ray lasing without population inversion in ^{3}He using Pauli principle. J Mod Opt 54:2607
54. Zibrov AS, Lukin MD, Nikonov DE, Hollberg L, Scully MO, Velichansky VL, Robinson HG (1995) Experimental demonstration of laser oscillation without population inversion via quantum interference in Rb. Phys Rev Lett 75:1499
55. Padmabandu GG, Welch GR, Shubin IN, Fry ES, Nikonov DE, Lukin MD, Scully MO (1996) Laser oscillation without population inversion in a sodium atomic beam. Phys Rev Lett 76:2053
56. Sete EA, Svidzinsky AA, Rostovtsev YV, Eleuch H, Jha PK, Suckewer S, Scully MO (2011) Using quantum coherence to generate gain in the XUV and X-ray: Gain-swept superradiance and lasing without inversion. IEEE J Sel Top Quantum Electron 18:541
57. Xia H, Svidzinsky AA, Yuan L, Lu S, Suckewer S, Scully MO (2012) Observing superradiant decay of excited-state helium atoms inside helium plasma. Phys Rev Lett 109:093604
58. Scully MO, Fleischhauer M (1994) Lasers without inversion. Science 263:337
59. Vagizov F, Antonov V, Radeonychev YV, Shakhmuratov RN, Kocharovskaya O (2014) Coherent Control of the waveforms of recoilless gamma-photons. Nature 508:80

Optical Communication: Its History and Recent Progress

Govind P. Agrawal

G.P. Agrawal (✉)
The Institute of Optics, University of Rochester, Rochester NY 14627, USA
e-mail: Govind.Agrawal@rochester.edu

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_8

8.1 Historical Perspective

The use of light for communication purposes dates back to antiquity if we interpret optical communication in a broad sense, implying any communication scheme that makes use of light. Most civilizations have used mirrors, fire beacons, or smoke signals to convey a single piece of information (such as victory in a war). For example, it is claimed that the Greeks constructed in 1084 B.C. a 500-km-long line of fire beacons to convey the news of the fall of Troy [1]. The chief limitation of such a scheme is that the information content is inherently limited and should be agreed upon in advance. Attempts were made throughout history to increase the amount of transmitted information. For example, the North American Indians changed the color of a smoke signal for this purpose. Similarly, shutters were used inside lighthouses to turn the beacon signal on and off at predetermined intervals. This idea is not too far from our modern schemes in which information is coded on the light emitted by a laser by modulating it at a high speed [2].

In spite of such clever schemes, the distance as well as the rate at which information could be transmitted using semaphore devices was quite limited even during the eighteenth century. A major advance occurred in 1792 when Claude Chappe came up with the idea of transmitting mechanically coded messages over long distances through the use of intermediate relay stations (10–15 km apart) that acted as *repeaters* in the modern-day language [3]. Figure 8.1 shows the inventor and his basic idea schematically. Chappe called his invention *optical telegraph* and developed a coding scheme shown in Fig. 8.1 to represent the entire alphabet through different positions of two needles. This allowed transmission of whole sentences over long distances. The first such optical telegraph was put in service in July 1794 between Paris and Lille (two French cities about 200 km apart). By 1830, the network had expanded throughout Europe [4]. The role of light in such systems was simply to make the coded signals visible so that they could be intercepted by the relay stations. The opto-mechanical

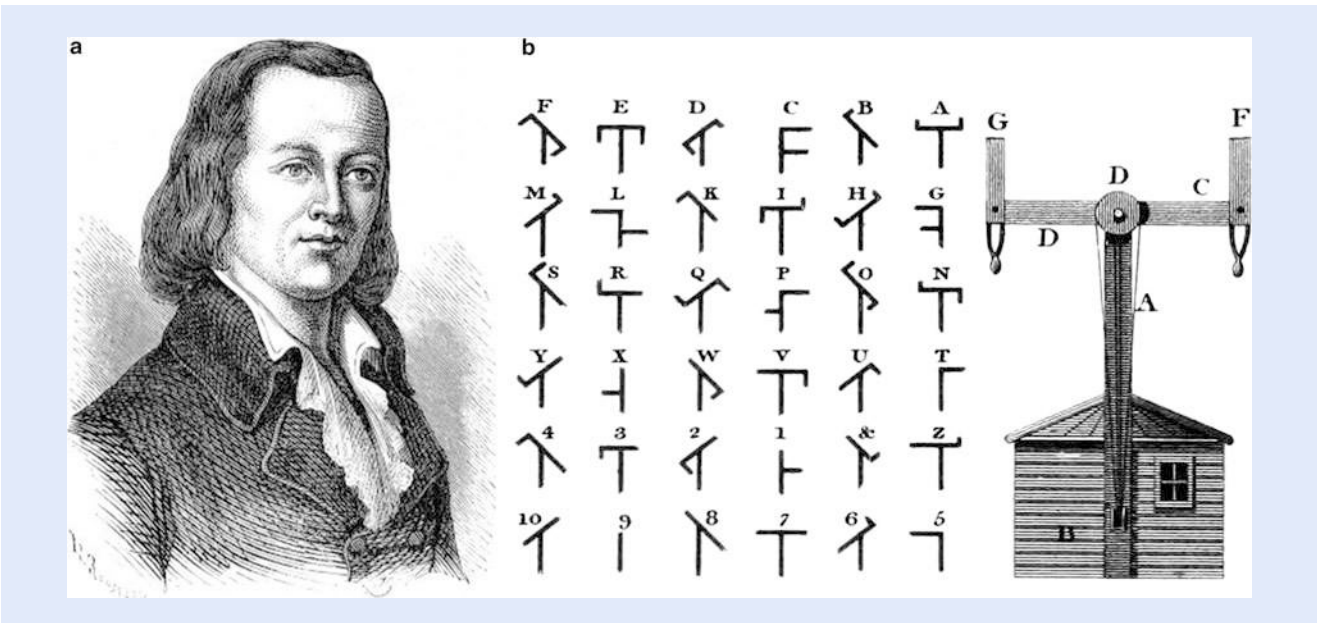


Fig. 8.1 Claude Chappe, his coding scheme, and the mechanical device used for making optical telegraphs (licensed under Public Domain via Wikimedia Commons)

communication systems of the nineteenth century were naturally slow. In modern-day terminology, the effective bit rate of such systems was less than 1 bit/s; a bit is the smallest unit of information in a binary system.

The advent of the electrical telegraph in the 1830s replaced the use of light by electricity and began the era of electrical communications [5]. The bit rate B could be increased to a few bits/s by using new coding techniques such as the Morse code. The use of intermediate relay stations allowed communication over long distances. Indeed, the first successful transatlantic telegraph cable went into operation in 1866. Telegraphy employed a digital scheme through two electrical pulses of different durations (dots and dashes of the Morse code). The invention of the telephone in 1876 brought a major change inasmuch as electric signals were transmitted in an analog form through a continuously varying electric current [6]. Analog electrical techniques dominated communication systems until the switch to optical schemes 100 years later.

The development of a worldwide telephone network during the twentieth century led to many advances in electrical communication systems. The use of coaxial cables in place of twisted wires increased system capacity considerably. The first coaxial-cable link, put into service in 1940, was a 3-MHz system capable of transmitting 300 voice channels (or a single television channel). The bandwidth of such systems was limited by cable losses, which increase rapidly for frequencies beyond 10 MHz. This limitation led to the development of microwave communication systems that employed electromagnetic waves at frequencies in the range of 1–10 GHz. The first microwave system operating at the carrier frequency of 4 GHz was put into service in 1948. Both the coaxial and microwave systems can operate at bit rates $\sim$ 100 Mbit/s. The most advanced coaxial system was put into service in 1975 and operated at a bit rate of 274 Mbit/s. A severe drawback of high-speed coaxial systems was their small repeater spacing ($\sim$ 1 km), requiring excessive regeneration of signals and making such systems expensive to operate. Microwave communication systems generally allowed for a larger repeater spacing but their bit rate was also limited to near 100 Mbit/s.

All of the preceding schemes are now classified under the general heading of telecommunication systems. A telecommunication system transmits information from one place to another, whether separated by a few kilometers or by transoceanic distances. It may but does not need to involve optics. The optical telegraph of Claude Chappe can be called the first optical telecommunication system that spread throughout Europe over a 40-year period from 1800 to 1840. However, it was soon eclipsed by electrical telecommunication systems based on telegraph and telephone lines. By 1950, scientists were again looking toward optics to provide solutions for enhancing the capacity of telecommunication systems. However, neither a coherent optical source nor a suitable transmission medium was available during the 1950s. The invention of the laser and its demonstration in 1960 solved the first problem. Attention was then focused on finding ways for using laser light for optical communication. Many ideas were advanced during the 1960s [7], the most noteworthy being the idea of light confinement using a sequence of gas lenses [8].

Optical fibers were available during the 1960s and were being used for making gastroscope and other devices that required only a short length of the fiber [9]. However, no one was serious about using them for optical communication. The main problem was that optical fibers available during the 1960s had such high losses that only 10 % of light entering at one end emerged from the other end of a fiber that was only a few meters long. Most engineers ignored them for telecommunication applications where light had to be transported over at least a few kilometers. It was suggested in 1966 that losses of optical fibers could be reduced drastically by removing impurities from silica glass used to make them, and that such low-losses fibers might be the best choice for optical

communication [10]. Indeed, Charles Kao was awarded one half of the 2009 noble prize for his groundbreaking achievements concerning the transmission of light in fibers for optical communication [11]. The idea of using glass fibers for optical communication was revolutionary since fibers are capable of guiding light in a manner similar to the confinement of electrons inside copper wires. As a result, they can be used in the same fashion as electric wires are used routinely.

However, before optical fibers could be used for optical communication, their losses had to be reduced to an acceptable level. This challenge was taken by Corning, an American company located not far from Rochester, New York where I work. A breakthrough occurred in 1970 when three Corning scientists published a paper indicating that they were able to reduce fiber losses to below 20 dB/km in the wavelength region near 630 nm [12]. Two years later, the same Corning team produced a fiber with a loss of only 4 dB/km by replacing titanium with germanium as a dopant inside the fiber's silica core. Soon after, many industrial laboratories entered the race for reducing fiber losses even further. The race was won in 1979 by a Japanese group that was able to reduce the loss of an optical fiber to near 0.2 dB/km in the infrared wavelength region near 1.55 μm [13]. This value was close to the fundamental limit set by the phenomenon of Rayleigh scattering. Even modern fibers exhibit loss values similar to those first reported in 1979.

In addition to low-loss optical fibers, switching from microwaves to optical waves also required a compact and efficient laser, whose output could be modulated to impose the information that needed to be transmitted over such fibers. The best type of laser for this purpose was the semiconductor laser. Fortunately, at about the same time Corning announced its low-loss fiber in 1970, GaAs semiconductor lasers, operating continuously at room temperature, were demonstrated by two groups working in Russia [14] and at Bell Laboratories [15]. The simultaneous availability of compact optical sources and low-loss optical fibers led to a worldwide effort for developing fiber-optic communication systems [16].

The first-generation systems were designed to operate at a bit rate of 45 Mbit/s in the near-infrared spectral region because GaAs semiconductor lasers used for making them emit light at wavelengths near 850 nm. Since the fiber loss at that wavelength was close to 3 dB/km, optical signal needed to be regenerated every 10 km or so using the so-called repeaters. This may sound like a major limitation, but it was better than the prevailing coaxial-cable technology that required regeneration every kilometer or so. Extensive laboratory development soon led to several successful field trials. AT&T sent its first test signals on April 1, 1977 in Chicago's Loop district. Three weeks later, General Telephone and Electronics sent live telephone traffic at 6 Mbit/s in Long Beach, California. It was followed by the British Post Office that began sending live telephone traffic through fibers near Martlesham Heath, UK. These trials were followed with further development, and commercial systems began to be installed in 1980. The new era of fiber-optic communication systems had finally arrived. Although not realized at that time, it was poised to revolutionize how humans lived and interacted. This became evident only after the advent of the Internet during the decade of the 1990s.

A commonly used figure of merit for communication systems is the bit rate–distance product BL, where B is the bit rate and L is the repeater spacing, the distance after which an optical signal must be regenerated to maintain its fidelity [2]. Figure 8.2 shows how the BL product has increased by a factor of 10^{18} through technological advances during the last 180 years. The acronym WDM in this figure stands for wavelength-division multiplexing, a technique used after 1992 to transmit multiple channels at different wavelengths through the same fiber. Its use enhanced the capacity of fiber-optic communication systems so dramatically that data transmission at 1 Tbit/s was realized by 1996. The acronym

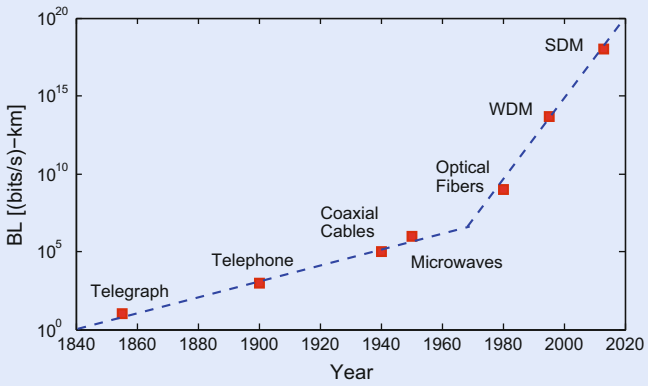


Fig. 8.2 Increase in the *BL* product during the period 1840–2015. The emergence of new technologies is marked by *red squares*. *Dashed line* shows the trend as an aid for the eye. Notice the change in slope around 1977 when optical fibers were first used for optical communications

SDM stands for space-division multiplexing, a technique used after 2010 to further enhance the capacity of fiber-optic systems in response to continuing increase in the Internet data traffic (with the advent of video streaming by companies such as YouTube and Netflix) and fundamental capacity limitations of single-mode fibers (see Sect. 8.5). Two features of Fig. 8.2 are noteworthy. First, a straight line in this figure indicates an exponential growth because of the use of logarithmic scale for the data plotted on the *y* axis. Second, a sudden change in the line's slope around 1977 indicates that the use of optical fibers accelerated the rate of exponential growth and signaled the emergence of a new optical communication era.

8.2 Basic Concepts Behind Optical Communication

Before describing the technologies used to advance the state of the art of fiber-optic communication systems, it is useful to look at the block diagram of a generic communication system in Fig. 8.3a. It consists of an optical transmitter and an optical receiver connected to the two ends of a communication channel that can be a coaxial cable (or simply air) for electric communication systems but takes the form of an optical fiber for all fiber-optic communication systems.

8.2.1 Optical Transmitters and Receivers

The role of optical transmitters is to convert the information available in an electrical form into an optical form, and to launch the resulting optical signal into a communication channel. Figure 8.3b shows the block diagram of an optical transmitter consisting of an optical source, a data modulator, and electronic circuitry used to derive them. Semiconductor lasers are commonly used as optical sources, although light-emitting diodes (LEDs) may also be used for some less-demanding applications. In both cases, the source output is in the form of an electromagnetic wave of constant amplitude. The role of the modulator is to impose the electrical data on this carrier wave by changing its amplitude, or phase, or both of them. In the case of some less-demanding applications, the current injected into a semiconductor laser itself is modulated directly, alleviating the need of an expensive modulator.

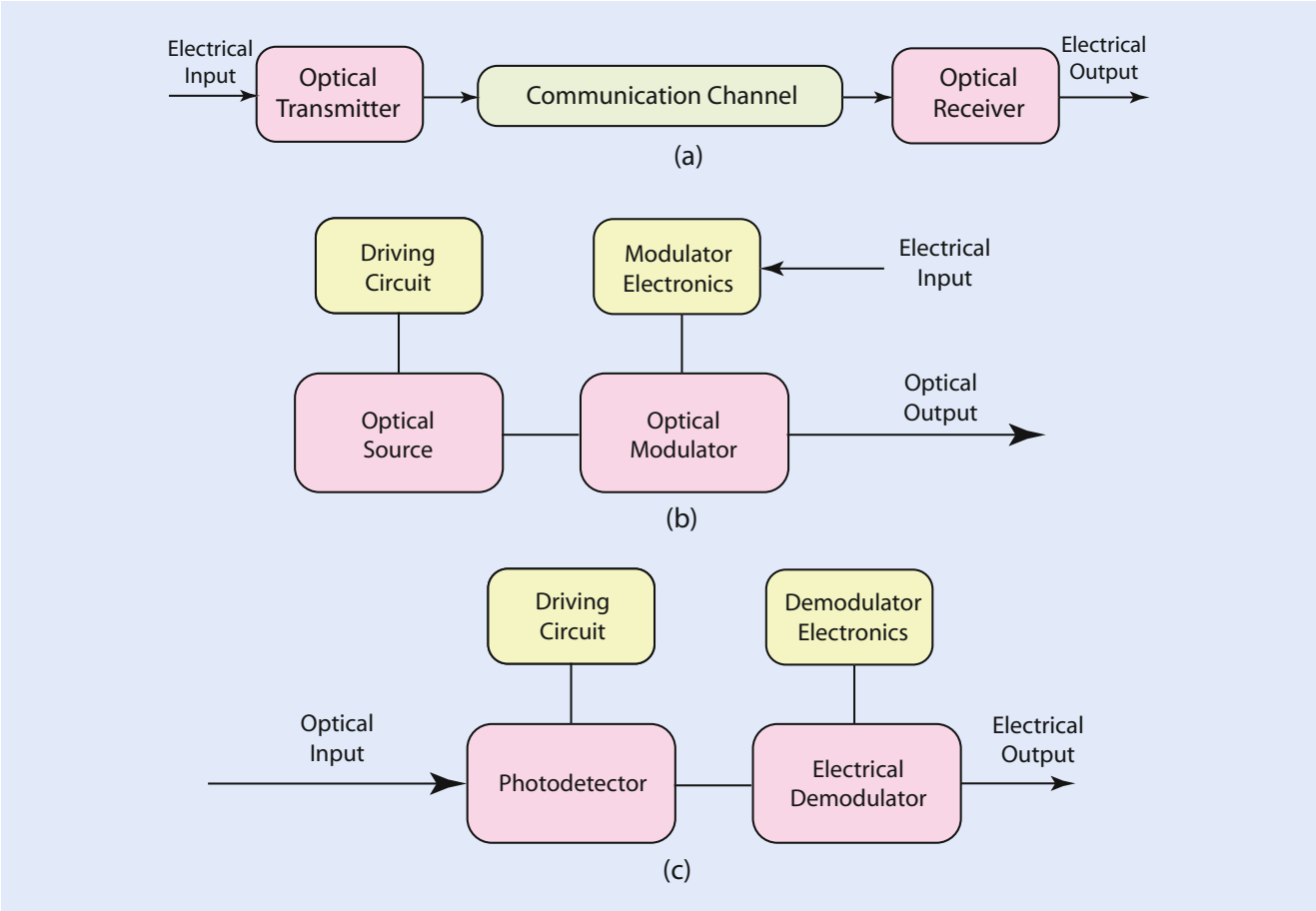


Fig. 8.3 (**a**) A generic optical communication system. (**b**) Components of an optical transmitter. (**c**) Components of an optical receiver

The role of optical receivers is to recover the original electrical data from the optical signal received at the output end of the communication channel. Figure 8.3c shows the block diagram of an optical receiver. It consists of a photodetector and a demodulator, together with the electronic circuitry used to derive them. Semiconductor photodiodes are used as detectors because of their compact size and low cost. The design of the demodulator depends on the modulation scheme used at the transmitter. Many optical communication systems employ a binary scheme referred to as intensity modulation with direct detection. Demodulation in this case is done by a decision circuit that identifies incoming bits as 1 or 0, depending on the amplitude of the electric signal. All optical receivers make some errors because of degradation of any optical signal during its transmission and detection, shot noise being the most fundamental source of noise. The performance of a digital lightwave system is characterized through the bit-error rate. It is customary to define it as the average probability of identifying a bit incorrectly. The error-correction codes are sometimes used to improve the raw bit-error rate of an optical communication system.

8.2.2 Optical Fibers and Cables

Most people are aware from listening to radios or watching televisions that electromagnetic waves can be transmitted through air. However, optical communication systems require electromagnetic waves whose frequencies lie in the visible

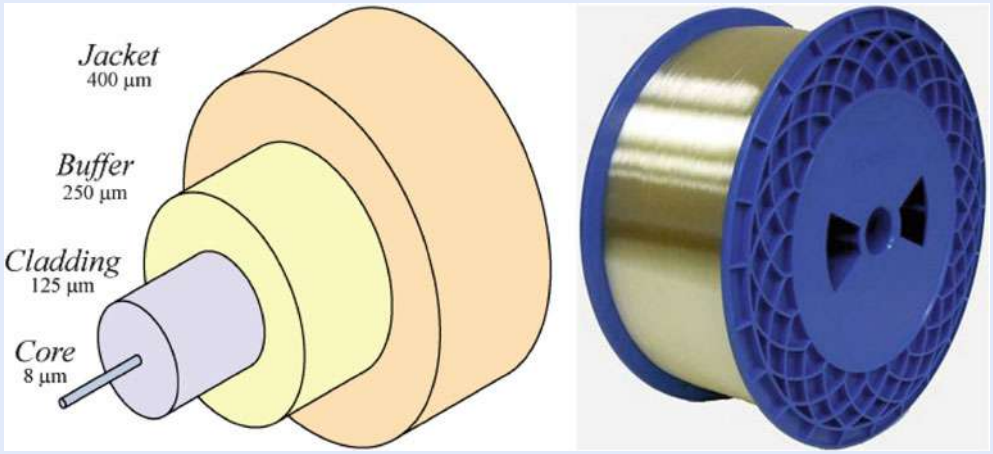


Fig. 8.4 Internal structure of a single-mode fiber; typical size of each part is indicated. A spool of such fiber is also shown (licensed under Public Domain via Wikimedia Commons)

or near-infrared region. Although such waves can propagate through air over short distances in good weather conditions, this approach is not suitable for making optical communication networks spanning the whole world. Optical fibers solve this problem and transmit light over long distances, irrespective of weather conditions, by confining the optical wave to the vicinity of a microscopic cylindrical glass core through a phenomenon known as *total internal reflection*.

Figure 8.4 shows the structure of an optical fiber designed to support a single spatial mode by reducing its core diameter to below 10 μm. In the case of a graded-index multimode fiber the core diameter is typically 50 μm. The core is made of silica glass and is doped with germania to enhance its refractive index slightly (by about 0.5 %) compared to the surrounding cladding that is also made of silica glass. A buffer layer is added on top of the cladding before putting a plastic jacket. The outer diameter of the entire structure, only a fraction of a millimeter, is so small that the fiber is barely visible. Before it can be used to transmit information, one or more optical fibers are enclosed inside a cable whose diameter may vary from 1 to 20 mm, depending on the intended application.

What happens to an optical signal transmitted through an optical fiber? Ideally, it should not be modified by the fiber at all. In practice, it becomes weaker because of unavoidable losses and is distorted through the phenomena such as chromatic dispersion and the Kerr nonlinearity [2]. As discussed earlier, losses were the limiting factor until 1970 when a fiber with manageable losses was first produced [12]. Losses were reduced further during the decade of 1970s, and by 1979 they have been reduced to a level as low as 0.2 dB/km at wavelengths near 1.55 μm. Figure 8.5 shows the wavelength dependence of power losses measured for such a fiber [13]. Multiple peaks in the experimental curve are due to the presence of residual water vapors. The dashed line, marked Rayleigh scattering, indicates that, beside water vapors, most of the loss can be attributed to the fundamental phenomenon of Rayleigh scattering, the same one responsible for the blue color of our sky. Indeed, although water peaks have nearly disappeared in modern fibers, their losses have not changed much as they are still limited by Rayleigh scattering.

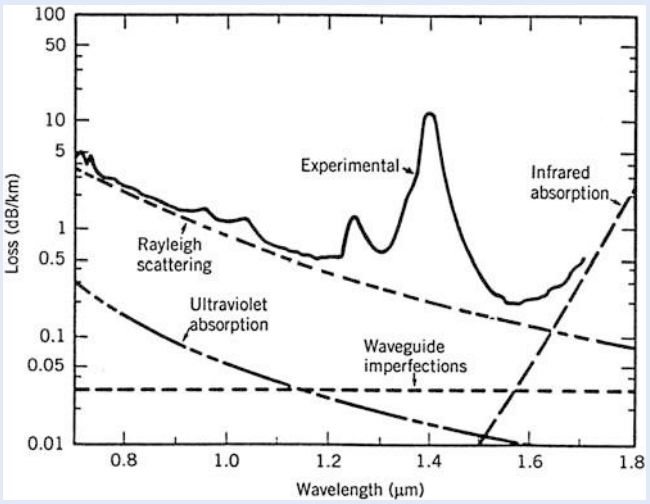


Fig. 8.5 Wavelength dependence of power losses measured in 1979 for a low-loss silica fiber [13]. Various lines show the contribution of different sources responsible for losses (from [2]; ©2010 Wiley)

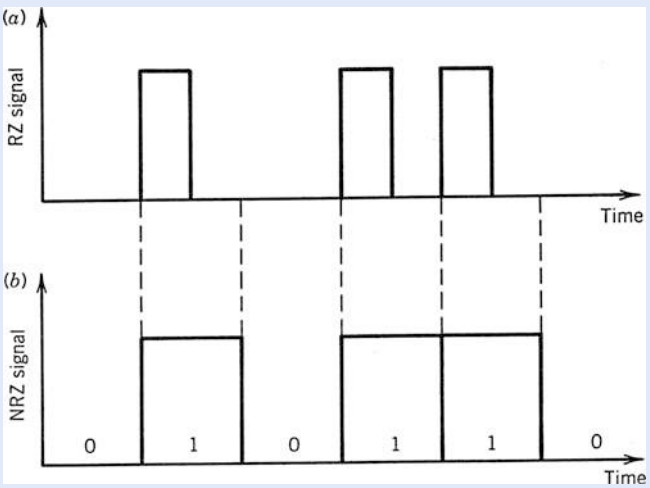


Fig. 8.6 Digital bit stream 010110. . . coded by using (**a**) return-to-zero (RZ) and (**b**) nonreturn-to-zero (NRZ) formats

8.2.3 Modulations Formats

The first step in the design of any optical communication system is to decide how the electrical binary data would be converted into an optical bit stream. As mentioned earlier, an electro-optic modulator is used for this purpose. The simplest technique employs optical pulses such that the presence of a pulse in the time slot of a bit corresponds to 1, and its absence indicates a 0 bit. This is referred to as on–off keying since the optical signal is either "off" or "on" depending on whether a 0 or 1 bit is being transmitted.

There are still two choices for the format of the resulting optical bit stream. These are shown in Fig. 8.6 and are known as the return-to-zero (RZ) and

nonreturn-to-zero (NRZ) formats. In the RZ format, each optical pulse representing bit 1 is shorter than the bit slot, and its amplitude returns to zero before the bit duration is over. In the NRZ format, the optical pulse remains on throughout the bit slot, and its amplitude does not drop to zero between two or more successive 1 bits. As a result, temporal width of pulses varies depending on the bit pattern, whereas it remains the same in the case of RZ format. An advantage of the NRZ format is that the bandwidth associated with the bit stream is smaller by about a factor of 2 simply because on–off transitions occur fewer times. Electrical communication systems employed the NRZ format for this reason in view of their limited bandwidth. The bandwidth of optical communication systems is large enough that the RZ format can be used without much concern. However, the NRZ format was employed initially. The switch to the RZ format was made only after 1999 when it was found that its use helps in designing high-capacity lightwave systems. By now, the RZ format is use almost exclusively for WDM systems whose individual channels are designed to operate at bit rates exceeding 10 Gbit/s.

8.2.4 Channel Multiplexing

Before the advent of the Internet, telephones were used most often for communicating information. When an analog electric signal representing human voice is digitized, the resulting digital signal contains 64,000 bits over each one second duration. The bit rate of such an optical bit stream is clearly 64 kbit/s. Since fiber-optic communication systems are capable of transmitting at bit rates of up to 40 Gbit/s, it would be a huge waste of bandwidth if a single telephone call was sent over an optical fiber. To utilize the system capacity fully, it is necessary to transmit many voice channels simultaneously through multiplexing. This can be accomplished through time-division multiplexing (TDM) or WDM. In the case of TDM, bits associated with different channels are interleaved in the time domain to form a composite bit stream. For example, the bit slot is about 15 μs for a single voice channel operating at 64 kb/s. Five such channels can be multiplexed through TDM if the bit streams of successive channels are delayed by 3 μs. Figure 8.7a shows the resulting bit stream schematically at a composite bit rate of 320 kb/s. In the case of WDM, the channels are spaced apart in the frequency domain. Each channel is carried by its own carrier wave. The carrier frequencies are spaced more than the channel bandwidth so that the channel spectra do not overlap, as seen in Fig. 8.7b. WDM is suitable for both analog and digital signals and is used in broadcasting of radio and television channels. TDM is readily implemented for digital signals and is commonly used for telecommunication networks.

The concept of TDM has been used to form digital hierarchies. In North America and Japan, the first level corresponds to multiplexing of 24 voice channels with a composite bit rate of 1.544 Mb/s (hierarchy DS-1), whereas in Europe 30 voice channels are multiplexed, resulting in a composite bit rate of 2.048 Mb/s. The bit rate of the multiplexed signal is slightly larger than the simple product of 64 kb/s with the number of channels because of extra control bits that are added for separating channels at the receiver end. The second-level hierarchy is obtained by multiplexing four DS-1 channels. This results in a bit rate of 6.312 Mb/s (hierarchy DS-2) for North America and 8.448 Mb/s for Europe. This procedure is continued to obtain higher-level hierarchies.

The lack of an international standard in the telecommunication industry during the 1980s led to the advent of a new standard, first called the synchronous

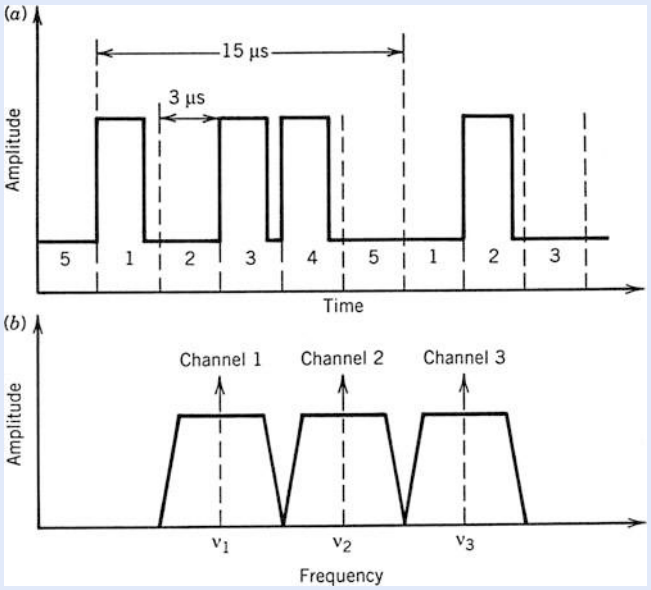


Fig. 8.7 (**a**) Time-division multiplexing of five digital voice channels operating at 64 kb/s. (**b**) Wavelength-division multiplexing of three analog or digital signals

Table 8.1 SONET/SDH bit rates

SONET	SDH	*B* (Mb/s)	Channels
OC-1		51.84	672
OC-3	STM-1	155.52	2016
OC-12	STM-4	622.08	8064
OC-48	STM-16	2488.32	32,256
OC-192	STM-64	9953.28	129,024
OC-768	STM-256	39,813.12	516,096

optical network (SONET) and later termed the synchronous digital hierarchy (SDH). It defines a synchronous frame structure for transmitting TDM digital signals. The basic building block of the SONET has a bit rate of 51.84 Mbit/s. The corresponding optical signal is referred to as OC-1, where OC stands for optical carrier. The basic building block of the SDH has a bit rate of 155.52 Mbit/s and is referred to as STM-1, where STM stands for a synchronous transport module. A useful feature of the SONET and SDH is that higher levels have a bit rate that is an exact multiple of the basic bit rate. Table 8.1 lists the correspondence between SONET and SDH bit rates for several levels. Commercial STM-256 (OC-768) systems operating near 40 Gbit/s became available by 2002. One such optical channel transmits more than half million telephone conversations over a single optical fiber. If the WDM technique is employed to transmit 100 channels at different wavelengths, one fiber can transport more than 50 million telephone conversations at the same time.

8.3 Evolution of Optical Communication from 1975 to 2000

As mentioned earlier, initial development of fiber-optic communication systems started around 1975. The enormous progress realized over the 40-year period extending from 1975 to 2015 can be grouped into several distinct generations. Figure 8.8 shows the increase in the *BL* product over the period 1975–2000 as quantified through various laboratory experiments [17]. The straight line corresponds to a doubling of the *BL* product every year. The first four generations of lightwave systems are indicated in Fig. 8.8. In every generation, the *BL* product increases initially but then begins to saturate as the technology matures. Each new generation brings a fundamental change that helps to improve the system performance further.

8.3.1 The First Three Generations

The first generation of optical communication systems employed inside their optical transmitters GaAs semiconductor lasers operating at a wavelength near 850 nm. The optical bit stream was transmitted through graded-index multimode fibers before reaching an optical receiver, where it was converted back to the electric domain using a silicon photodetector. After several field trials during the period 1977–1979, such systems became available commercially in the year 1980. They operated at a bit rate of 45 Mbit/s and allowed repeater spacings of up to 10 km. The larger repeater spacing compared with 1-km spacing of coaxial systems was an important motivation for system designers because it decreased the installation and maintenance costs associated with each repeater. It is important to stress that even the first-generation systems transmitted nearly 700 telephone calls simultaneously over a single fiber through the use of TDM.

It was evident to system designers that the repeater spacing could be increased considerably by operating the system in the infrared region near 1.3 μm, where fiber losses were below 1 dB/km (see Fig. 8.5). Furthermore, optical fibers exhibit minimum dispersion in this wavelength region. This realization led to a worldwide effort for the development of new semiconductor lasers and detectors based on the InP material and operating near 1.3 μm. The second generation of

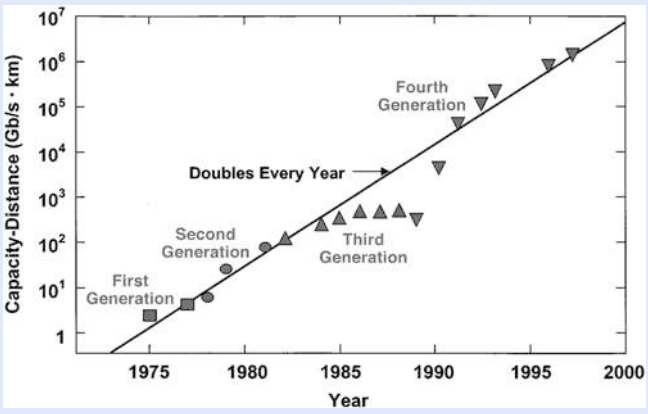


Fig. 8.8 Increase in the *BL* product over the period 1975–2000 through several generations of optical communication systems. Different symbols are used for successive generations (from [2]; ©2010 Wiley)

fiber-optic communication systems became available in the early 1980s, but their bit rate was initially limited to 100 Mbit/s because of dispersion in multimode fibers. This limitation was overcome by the use of single-mode fibers. In such fibers the core diameter is reduced to near 10 μm (see Fig. 8.4) so that the fiber supports a single spatial mode. A laboratory experiment in 1981 demonstrated transmission at 2 Gbit/s over 44 km of a single-mode fiber. The introduction of commercial systems soon followed. By 1987, such second-generation commercial systems were operating at bit rates of up to 1.7 Gbit/s with a repeater spacing of about 50 km.

The repeater spacing of the second-generation systems was still limited by the fiber loss at their operating wavelength of 1.3 μm. As seen in Fig. 8.5, losses of silica fibers become the smallest at wavelengths near 1.55 μm. However, the introduction of third-generation optical communication systems operating at 1.55 μm was considerably delayed by a relatively large dispersion of single-mode fibers in this spectral region. Conventional InGaAsP semiconductor lasers could not be used because of pulse spreading occurring as a result of simultaneous oscillation of multiple longitudinal modes. This dispersion problem could be solved either by using dispersion-shifted fibers designed to have minimum dispersion near 1.55 μm, or by designing lasers such that their spectrum contained a dominant single longitudinal mode. Both approaches were followed during the 1980s. By 1985, laboratory experiments indicated the possibility of transmitting information at bit rates of up to 4 Gbit/s over distances in excess of 100 km. Third-generation optical communication systems operating at 2.5 Gbit/s became available commercially in 1990, and their bit rate was soon extended to 10 Gbit/s. The best performance was achieved using dispersion-shifted fibers in combination with lasers oscillating in a single longitudinal mode.

A relatively large repeater spacing of the third-generation 1.55-μm systems reduced the need of signal regeneration considerably. However, economic pressures demanded further increase in its value of close to 100 km. It was not immediately obvious how to proceed since losses of silica fibers at 1.55 μm were limited to near 0.2 dB/km by the fundamental process of Rayleigh scattering. One solution was to develop more sensitive optical receivers that could work reliably at reduced power levels. It was realized by many scientists that repeater spacing could be increased by making use of a heterodyne-detection scheme (similar to that used for radio- and microwaves) because its use would require less power at the optical receiver. Such systems were referred to as coherent lightwave systems and were under development worldwide during the 1980s. However, the deployment of such systems was postponed with the advent of fiber amplifiers in 1989 that were pumped optically using semiconductor lasers and were capable of boosting the signal power by a factor of several hundreds.

8.3.2 The Fourth Generation

By 1990 the attention of system designers shifted toward using three new ideas: (1) periodic optical amplification for managing fiber losses, (2) periodic dispersion compensation for managing fiber dispersion, and (3) WDM for enhancing the system capacity. As seen in Fig. 8.9, the WDM technique employs multiple lasers at slightly different wavelengths such that multiple data streams are transmitted simultaneously over the same optical fiber. The basic idea of WDM was not new as this technique was already being used at the radio- and microwave frequencies (e.g., by the television industry). However, its adoption at optical wavelengths required the development of a large number of new devices over a time span of a few years. For example, optical multiplexers and demultiplexers that

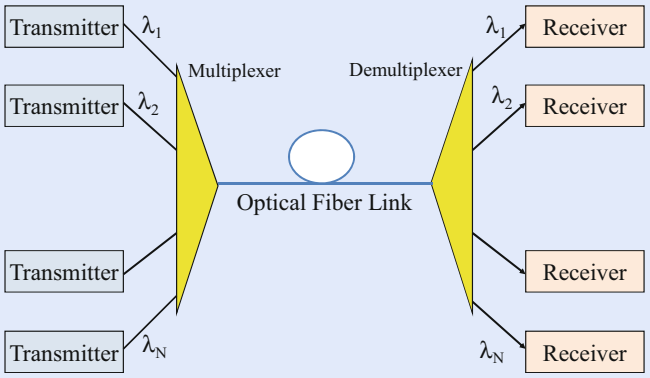


Fig. 8.9 Schematic of a WDM communication system. Multiple transmitters operating at different wavelengths are combined using a multiplexer and all channels are sent simultaneously over the same optical fiber. A demultiplexer at the receiving end separates individual channels and sends them to different receivers

could combine and separate individual channels at the two ends of a fiber link were critical components for the advent of the WDM systems.

The fourth generation of optical communication systems made use of optical amplifiers for increasing the repeater spacing, in combination with the WDM technique for increasing the system capacity. As seen in Fig. 8.8, the advent of the WDM technique around 1992 started a revolution that resulted in doubling of the system capacity every 8 months or so and led to lightwave systems operating at a bit rate of 1 Tbit/s by 1996. In most WDM systems, fiber losses are compensated periodically using erbium-doped fiber amplifiers spaced 60–80 km apart. Such WDM systems operating at bit rates of up to 80 Gbit/s were available commercially by the end of 1995. The researchers worldwide were pushing the limit of WDM technology. By 1996, three research groups reported during a post-deadline session of the Optical Fiber Communications conference that they were able to operate WDM systems with the total capacity of more than 1 Tbit/s. This represented an increase in the system capacity by a factor of 400 over a period of just 6 years!

The emphasis of most WDM systems is on transmitting as many optical channels as possible over a single fiber by adding more and more lasers operating at different wavelengths. The frequency spacing between two neighboring channels is chosen to be as small as possible but it has to be larger than the bandwidth of each channel. At a bit rate of 40 Gbit/s, a channel spacing of 50 GHz is the smallest that can be used. The standard setting agency, ITU, has assigned a set of fixed frequencies for commercial WDM systems using this 50-GHz channel spacing. All these frequencies lie in the wavelength region near 1550 nm where fiber losses are the smallest. The wavelength range of 1530–1570 nm is called the C band (C standing for conventional), and most commercial WDM systems are designed to work in this band. However, the S and L bands lying on the short- and long-wavelength side of the C band, respectively, are also used if necessary. This approach led in 2001 to a 11-Tbit/s experiment in which 273 channels, each operating at 40 Gbit/s, were transmitted over a distance of 117 km [18]. Given that the first-generation systems had a capacity of 45 Mbit/s in 1980, it is remarkable that the use of WDM increased the system capacity by a factor of more than 200,000 over a period of 21 years.

8.3.3 Bursting of the Telecom Bubble in 2000

It should be clear from Fig. 8.8 and the preceding discussion that the adoption of WDM during the fourth generation of optical communication systems was a disruptive technology. Fortunately, its adoption coincided with the advent and commercialization of the Internet around 1994. Just as the explosive growth of websites all around the world increased the volume of data flowing through the telecommunication networks, the fourth generation of optical communication systems making use of the WDM technology became available. Its advent allowed telecom operators to manage the data traffic by simply adding more channels to an existing WDM system. The demand went through the roof during the 4-year span of 1996–2000, and it led to a stock-market bubble that is now referred to as the telecom bubble.

The formation of the telecom bubble was the result of a rapid growth after 1995 in the telecommunication business. The stocks of companies dealing with the manufacturing and delivery of telecom services soared after 1995. As an example, consider the company JDS–Fitel involved in selling various optical devices needed for telecom systems. Its stock value was around $8 in June 1994, jumped to near $20 in June 1995, and exceeded $70 in June 1996 when the stock was split by 2:1 to bring the price near $35. The stock was split again in November 1997 when its price doubled a second time. In early 1999 the company announced a merger with Uniphase, another fast-growing optics company, resulting in the formation of JDSU. During that year, the stock of JDSU was increasing so rapidly that it was split two more times. Figure 8.10 shows how the stock price of JDSU varied over the 8-year period ranging from 1996 to 2004 after taking into account multiple splits. A nearly exponential growth during the 1999 indicates the formation of the telecom bubble during that year. A similar growth occurred in the stock price of many other telecommunication and Internet companies.

The telecom bubble burst during the year 2000, and the stock prices of all telecommunication companies collapsed soon after, including that of JDSU as seen in Fig. 8.10. Several companies went out of business and many surviving were in trouble financially. Commercial WDM systems capable of operating at 1 Tbit/s were still being sold, but their was no buyer for them. The research and development of fiber-optic communications systems slowed down to a crawl as everyone waited for the revival of the telecom industry. It took nearly 5 years before the US economy recovered, only to crash again in August 2008 owing to the formation of another bubble, this time in the real-estate market. One can say that the decade of 2000–2009 has not been a kind one as far as the telecommunication industry is concerned.

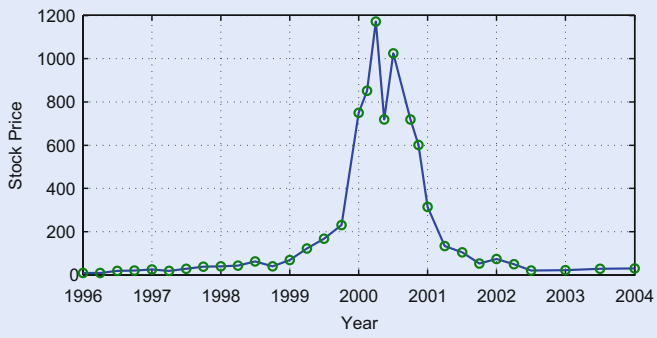


Fig. 8.10 Price of JDSU stock over a period extending from January 1996 to January 2004. A sharp rise in price during the year 1999 is followed by a sharp decline after July 2001 (*source*: public domain data)

8.4 The Fifth Generation

In spite of the two severe economical downturns, considerable progress has occurred since 2000 in designing advanced optical communication systems, leading to the fifth and sixth generations of such systems. The focus of fifth-generation systems was on making the WDM systems more efficient spectrally. This was accomplished by reviving the coherent detection scheme that was studied in the late 1980s but abandoned soon after fiber-based optical amplifiers became available. Coherent receivers capable of detecting both the amplitude and phase of an optical signal through a heterodyne scheme were developed soon after the year 2000. Their commercial availability near the end of the decade allowed system designers to employ advanced modulation formats in which information is encoded using both the amplitude and phase of an optical carrier.

The basic concept can be understood from Fig. 8.11, showing four modulation formats using the so-called constellation diagram that displays the real and imaginary parts of the complex electric field along the x and y axes, respectively. The first configuration represents the standard binary format, called amplitude-shift keying (ASK), in which the amplitude or intensity of the electric field takes two values, marked by circles and representing 0 and 1 bits of a digital signal. The second configuration is another binary format, called phase-shift keying (PSK), in which the amplitude remains constant but phase of the electric field takes two values, say 0 and π, that represent the 0 and 1 bits of a digital signal. The third configuration in part (c) of Fig. 8.11 shows the quaternary PSK (or QPSK) format in which the optical phase takes four possible values. This case allows to reduce the signal bandwidth since two bits can be transmitted during each time slot, and the effective bit rate is halved. Borrowing from microwave communication terminology, the reduced bit rate is called the symbol rate (or baud). The last example in Fig. 8.11 shows how the symbol concept can be extended to multilevel signaling such that each symbol carries 4 bits or more. An additional factor of two can be gained if one transmits two orthogonally polarized symbols simultaneously during each symbol slot, a technique referred to as polarization division multiplexing.

The concept of spectral efficiency, defined as the number of bits transmitted in 1 s within a 1-Hz bandwidth, is quite useful in understanding the impact of coherent detection in combination with phase-encoded modulation formats. The spectral efficiency of fourth generation WDM systems that employed ASK as the modulation format was limited to below 0.8 bit/s/Hz since at most 40 billion bits/s could be transmitted over a 50-GHz bandwidth of each WDM channel. This value for the fifth-generation systems can easily exceed 3 by using polarization multiplexing in combination with the QPSK format. Figure 8.12 shows how the spectral efficiency of optical communication systems has evolved since 1990 when its value was near 0.05 bit/s/Hz. Values near 2 bit/s/Hz were realized by 2005 and they approached 10 bit/s/Hz by the year 2010 [19].

The availability of coherent receivers and increased computing speeds led to another advance after it was realized that one can use digital signal processing to

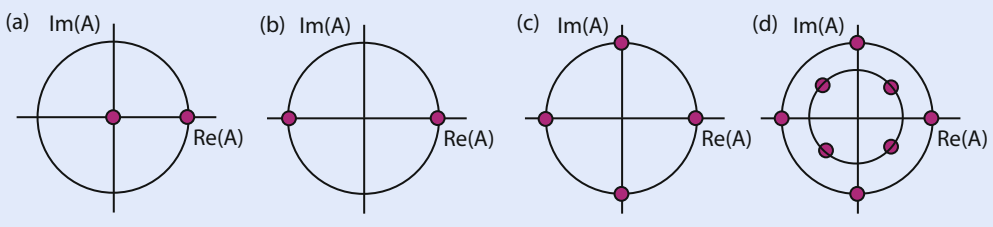


Fig. 8.11 Constellation diagrams for (**a**) ASK, (**b**) PSK, (**c**) QPSK, and (**d**) multilevel QPSK formats

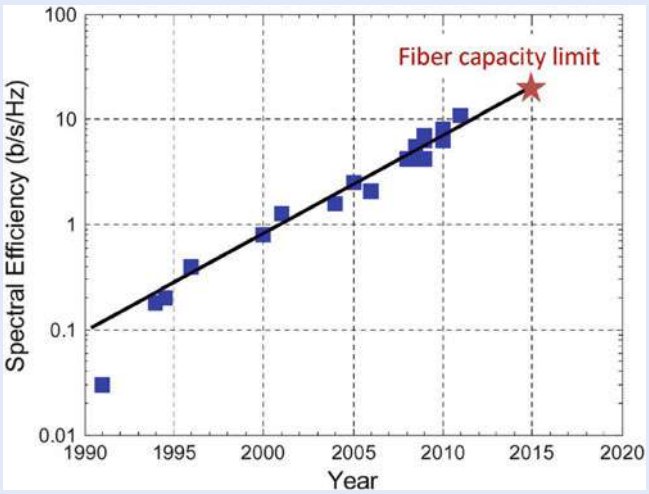


Fig. 8.12 Evolution of spectral efficiency after 1990 through laboratory demonstrations. The *red star* shows the fundamental capacity limit of optical fibers (after [19]; ©2012 IEEE)

improve the signal-to-noise ratio (SNR) of an optical signal arriving at the receiver. Since a coherent receiver detects both the amplitude and the phase of an optical signal, together with its state of polarization, one has in essence a digital representation of the electric field associated with the optical signal. As a result, special electronic chips can be designed to process this digital signal that can compensate for the degradation caused by such unavoidable factors as fiber dispersion. One can also implement error-correcting codes and employ encoder and decoder chips to improve the bit-error rate at the receiver end. A new record was set in 2011 when 64-Tbit/s transmission was realized over 320 km of a single-mode fiber using 640 WDM channels that spanned both the C and L bands with 12.5-GHz channel spacing [20]. Each channel contained two polarization-multiplexed 107-Gbit/s signals coded with a modulation format known as quadrature amplitude modulation. Such techniques are routinely implemented in modern optical communication systems.

8.5 The Sixth Generation

As the capacity of WDM systems approached 10 Tbit/s, indicating that 10 trillion bits could be transmitted each second over a single piece of optical fiber supporting a single optical mode inside its tiny core (diameter about 10 μm), scientists began to think about the ultimate information capacity of a single-mode fiber. The concept of the channel capacity C was first introduced by Shannon in a 1948 paper [21] in which he showed that the SNR sets the fundamental limit for any linear communication channel with a finite bandwidth W through the remarkably simple relation $C = W\log_2(1 + SNR)$. The spectral efficiency, defined as $SE = C/W$, is thus only limited by the SNR of the received signal and can, in principle, be increased indefinitely by sending more and more powerful signals over the channel. Unfortunately, this conclusion does not hold for optical fibers that are inherently nonlinear and affect the bit stream propagating through them in a nonlinear fashion [2].

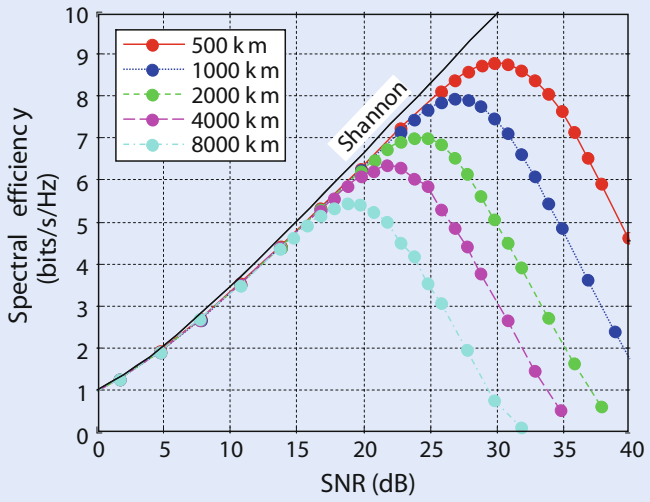


Fig. 8.13 Spectral efficiency as a function of SNR calculated numerically including the nonlinear effects over transmission distances ranging from 500 to 8000 km (after [22]; ©2010 IEEE)

8.5.1 Capacity Limit of Single-Mode Fibers

Considerable attention was paid during the decade of 2000 to estimating the ultimate capacity of single-mode fibers in the presence of various nonlinear effects. In a paper published in 2010 Essiambre et al. were able to develop a general formalism for calculating it [22]. Figure 8.13 shows how the nonlinear effects reduce the spectral efficiency from its value predicted by Shannon's relation, when high signal powers are launched to ensure a high SNR at the receiver. As one may expect, the spectral efficiency depends on the transmission distance, and it becomes worse as this distance increases. However, the most noteworthy feature of Fig. 8.13 is that, for any transmission distance, spectral efficiency is maximum at an optimum value of SNR that changes with distance. For example, the spectral efficiency of a 1000-km-long link is limited to 8 bit/s/Hz for single polarization, irrespective of the modulation format employed. This is in sharp contrast to the prediction of Shannon and reflects a fundamental limitation imposed by the nonlinear effects.

We can use the results shown in Fig. 8.13 to estimate the ultimate capacity of a single-mode fiber. The usable bandwidth of silica fibers in the low-loss window centered around 1550 nm is about 100 nm. This value translates into a channel bandwidth of 12.5 THz. Using this value and a peak spectral efficiency of about 16 bit/s/Hz (assuming polarization-division multiplexing), the maximum capacity of a single-mode fiber is estimated to be 200 Tb/s. This is an enormous number and was thought to be high enough until recently that system designers did not worry about running out of capacity. However, data traffic over fiber-optic networks has experienced a steady growth, doubling every 18 months, since the advent of the Internet in the early 1990s. The growth has even accelerated in recent years owing to the new activities such as video streaming. One way to meet the demand would be to deploy more and more fiber cables. However, this approach will result in a larger and larger fraction of the total electrical power being devoted to supporting optical transport networks. It is estimated that by 2025 the energy demand of modern telecommunication systems will consume a very large fraction of the total US energy budget, unless a way is found to design energy efficient optical networks.

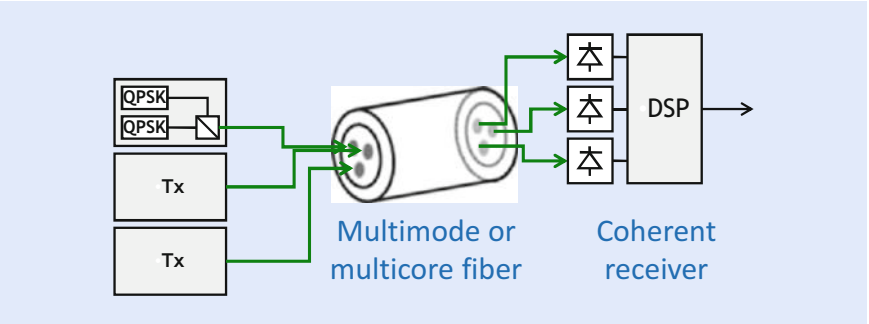


Fig. 8.14 Schematic illustration of the basic idea behind the SDM technique. WDM signals from different transmitters enter different cores or modes of a multimode fiber and are processed at the other end by different coherent receivers; DSP stands for digital signal processing (courtesy of S. Mumtaz)

8

8.5.2 Space-Division Multiplexing

One proposed solution makes use of space-division multiplexing (SDM) to increase the capacity of fiber-optic communication networks at a reduced energy cost per transmitted bit [23–26]. The basic idea is to employ multimode fibers such that several WDM bit streams can be transmitted over different modes of the same fiber. The energy advantage comes from integrating the functionalities of key optical components into a smaller number of devices. For instance, if a single multimode optical amplifier is used to amplify all spatially multiplexed bit streams, power consumption is likely to be lower compared to using separate amplifiers. For this reason, the SDM technique is attracting increasing attention since 2010, and several record-setting experiments have already been performed. Most of them employ multicore fibers in which several cores share the same cladding. Each core is typically designed to support a single mode but that is not a requirement. Figure 8.14 shows schematically the basic idea behind SDM using the case of a three-core fiber as an example.

Similar to the case of WDM technology, the implementation of SDM requires not only new types of fibers but also many other active and passive optical components such as mode multiplexers/demultiplexers and fiber amplifiers that can amplify signals in all modes/cores simultaneously. A lot of progress has been made since 2010 in realizing such devices and many laboratory demonstrations have shown the potential of SDM for enhancing the system capacity [23–26]. Figure 8.15 shows how the capacity of optical communication systems has evolved over a period ranging from 1980 to 2015 and covering all six generations. Single-wavelength systems, employing TDM in the electrical domain, started with a capacity of under 100 Mbit/s in the 1980s and were operating at 10 Gb/s around 1990. The advent of WDM in the early 1990 led to a big jump in the system capacity and subsequent adoption of coherent detection with digital signal processing allowed the capacity to reach 64 Tbit/s by the year 2010 [20]. Further increase in system capacity required the adoption of SDM. In a 2012 experiment, SDM was used to demonstrate data transmission at 1000 Tbit/s (or 1 Pbit/s) by employing a 12-core fiber [24]. Each fiber core carried 222 WDM channels, and each wavelength transmitted a 380-Gbit/s bit stream over a 52-km-long multicore fiber with a spectral efficiency of 7.6 bit/s/Hz.

The simplest SDM case corresponds to a multicore fiber whose cores are far enough apart that they experience little coupling. In this situation, WDM signals in each core travel independently, and the situation is analogous to using separate fibers. Indeed, most high-capacity experiments have employed this configuration

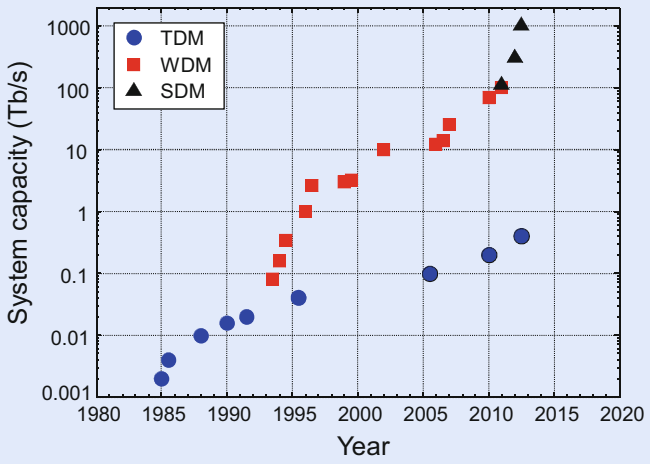


Fig. 8.15 Increase in the capacity of optical communication systems (on a logarithmic scale) realized from 1980 to 2015 using three different multiplexing techniques. Note the change in the slope around 1995 and 2011 when the WDM and SDM techniques were adopted (courtesy of R.J. Essiambre)

through multicore fibers with 7, 12, or 19 cores. In a second category of experiments single-core fibers supporting a few spatial modes are employed [23]. In this case, modes become invariably coupled, both linearly and nonlinearly, since all channels share the same physical path. Degradations induced by linear coupling are then removed at the receiver end through digital signal processing. In a 2015 experiment, a fiber supporting 15 spatial modes was used to transmit 30 - polarization-multiplexed channels over 23 km [27].

8.6 Worldwide Fiber-Optic Communication Network

The advent of the Internet in the early 1990s made it necessary to develop a worldwide network capable of connecting all computers (including cell phones) in a transparent manner. Such a network required deployment of fiber-based submarine cables across all oceans. The first such cable was installed in 1988 across the Atlantic ocean (TAT–8) but it was designed to operate at only 280 Mbit/s using the second-generation technology. The same technology was used for the first transpacific fiber-optic cable (TPC–3), which became operational in 1989. By 1990 the third-generation lightwave systems had been developed. The TAT–9 submarine system used this technology in 1991; it was designed to operate near 1.55 μm at a bit rate of 560 Mb/s with a repeater spacing of about 80 km. The increasing traffic across the Atlantic Ocean led to the deployment of the TAT–10 and TAT–11 cables by 1993 with the same technology. A submarine cable should be strong so that it can withstand biting by large sea animals. Figure 8.16 shows, as an example, the internal structure of a submarine cable containing several fibers for carrying bidirectional traffic. Optical fibers are immersed in a water-resistant jelly that is surrounded with many steel rods to provide strength. Steel rods are kept inside a copper tube that itself is covered with the insulating polyethylene. As the scale on the right side of Fig. 8.16 shows, the outer diameter of the entire cable is still only 1.7 cm.

After 1990, many laboratory experiments investigated whether the amplifier technology could be deployed for submarine cables such that the signal retained its optical form all along the cable length, thus avoiding the use of expensive in-line

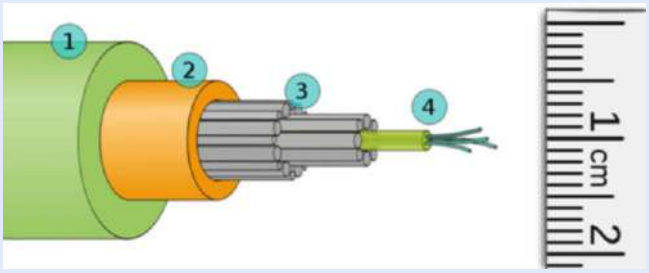


Fig. 8.16 Internal structure of a submarine cable containing several fibers: (1) insulating polyethylene; (2) copper tubing; (3) steel rods; (4) optical fibers in water-resistant jelly. The scale on the right shows the actual size (licensed under Public Domain via Wikimedia Commons)

Table 8.2 High-capacity submarine fiber-optic systems

System name	Year	Capacity (Tb/s)	Length (km)	WDM channels	Fiber pairs
VSNL transatlantic	2001	2.56	13,000	64	4
FLAG	2001	4.8	28,000	60	8
Apollo	2003	3.2	13,000	80	4
SEA-ME-WE 4	2005	1.28	18,800	64	2
Asia–America Gateway	2009	2.88	20,000	96	3
India-ME-WE	2009	3.84	13,000	96	4
African Coast to Europe	2012	5.12	13,000	128	4
West Africa Cable System	2012	5.12	14,500	128	4
Arctic fiber	2015	8.0	18,000	50	4

regenerators. As early as 1991, an experiment employed a recirculating-loop configuration to demonstrate the possibility of data transmission in this manner over 14,300 km at 5 Gbit/s. This experiment indicated that an all-optical, submarine transmission system was feasible for intercontinental communication. The TAT–12 cable, installed in 1995, employed optical amplifiers in place of in-line regenerators and operated at a bit rate of 5.3 Gbit/s with an amplifier spacing of about 50 km. The actual bit rate was slightly larger than the data rate of 5 Gbit/s because of the overhead associated with the forward-error correction that was necessary for the system to work. The design of such lightwave systems becomes quite complex because of the cumulative effects of fiber dispersion and nonlinearity, which must be controlled over long distances.

The use of the WDM technique after 1996 in combination with optical amplifiers, dispersion management, and error correction revolutionized the design of submarine fiber-optic systems. In 1998, a submarine cable known as AC–1 was deployed across the Atlantic Ocean with a capacity of 80 Gb/s using the WDM technology. An identically designed system (PC–1) crossed the Pacific Ocean. The use of dense WDM, in combination with multiple fiber pairs per cable, resulted in systems with large capacities. After 2000, several submarine systems with a capacity of more than 1 Tbit/s became operational (see Table 8.2). Figure 8.17 shows the international submarine cable network of fiber-optic communication systems. The VSNL transatlantic submarine system installed in 2001 had a total capacity of 2.56 Tbit/s and spans a total distance of 13,000 km. A submarine

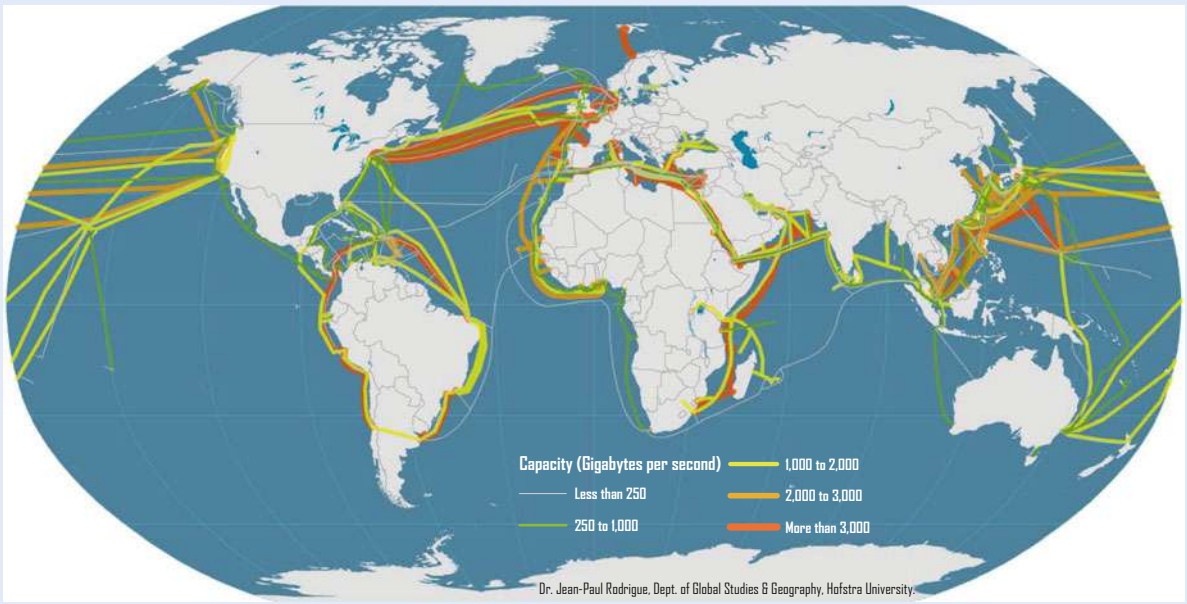


Fig. 8.17 International submarine cable network of fiber-optic communication systems around 2015 (*source*: dataset encoded by Greg Mahlknecht, ► http://www.cablemap.info)

system, known as India-ME-WE and installed in 2009, is capable of operating bidirectionally at 3.84 Tb/s through four fiber pairs. By 2012, submarine systems with a total capacity of 5 Tbit/s for traffic in each direction became operational. A proposed system, known as Arctic Fiber, will be capable of carrying traffic at speeds of up to 8 Tbit/s by transmitting 50 channels (each operating at 40 Gb/s) over four fiber pairs. It is estimated that more than 400 million kilometers of optical fiber have already been deployed worldwide, a number that is close to three times the distance to sun.

8.7 Conclusions

This chapter began with a brief history of optical communication before describing the main components of a modern optical communication system. Specific attention was paid to the development of low-loss optical fibers as they played an essential role after 1975. I describe in detail the evolution of fiber-optic communication systems through its six generations over a 40-year time period ranging from 1975 to 2015. I also discuss how the adoption of WDM during the 1990s was fueled by the advent of the Internet and how it eventually led in 2000 to bursting of the telecom bubble in the stock markets worldwide. However, the telecommunication industry recovered by 2005, and the researchers have come up with new techniques during the last 10 years. Recent advances brought by digital coherent technology and space-division multiplexing are described briefly in this chapter. Figure 8.17 shows the international submarine cable network of fiber-optic communication systems that allows the Internet to operate transparently, interconnecting computers worldwide on demand. Such a global high-speed network would have not been possible without the development of fiber-optic communication technology during the 1980s and the adoption of the WDM technique during the decade of the 1990s. One can only wonder what the future holds, especially if the potential of the SDM technology is realized by the year 2020.

Acknowledgements The author thanks Dr. R.-J. Essiambre of Alcatel–Lucent Bell Laboratories for helpful suggestions. The financial support of US National Science Foundation is also gratefully acknowledged.

References

1. Huurdeman AA (2003) The worldwide history of telecommunications. Wiley, Hoboken, NJ
2. Agrawal GP (2010) Fiber-optic communication systems, 4th edn. Wiley, Hoboken, NJ
3. Chappe I. (1824) Histoire de la télégraphie (in French), University of Michigan Library
4. Holzmann GJ, Pehrson B (2003) The early history of data networks. Wiley, Hoboken, NJ
5. Jones A (1852) Historical sketch of the electrical telegraph. Putnam, New York
6. Bell AG (1876) Improvement in telegraphy, U.S. Patent No. 174,465
7. Pratt WK (1969) Laser communication systems. Wiley, Hoboken, NJ
8. Miller SE (1966) Communication by laser. Sci Am 214:19–25
9. Hecht J (1999) City of light: the story of fiber optics. Oxford University Press, New York
10. Kao KC, Hockham GA (1966) Dielectric-fiber surface waveguides for optical frequencies. Proc IEE 113:1151–1158
11. ► http://www.nobelprize.org/nobel_prizes/physics/laureates/2009/
12. Kapron FP, Keck DB, Maurer RD (1970) Radiation losses in glass optical waveguides. Appl Phys Lett 17:423–425
13. Miya T, Terunuma Y, Hosaka T, Miyoshita T (1979) Ultimate low-loss single-mode fiber at 1.55 μm. Electron Lett 15:106–108
14. Alferov Z. (2000) Double heterostructure lasers: early days and future perspectives. IEEE J Sel Top Quant Electron 6:832–840
15. Hayashi I, Panish MB, Foy PW, Sumski S (1970). Junction lasers which operate continuously at room temperature. Appl Phys Lett 17:109–111
16. Willner AE (ed.) (2000) Several historical articles in this millennium issue cover the development of lasers and optical fibers. IEEE J Sel Top Quant Electron 6:827–1513
17. Kogelnik H (2000) High-capacity optical communications: personal recollections. IEEE J Sel Top Quant Electron 6:1279–1286
18. Fukuchi K, Kasamatsu T, Morie M, Ohhira R, Ito T, Sekiya K, Ogasahara D, Ono T (2001) 10.92-Tb/s (273 × 40-Gb/s) triple-band/ultra-dense WDM optical-repeatered transmission experiment. In: Proceedings of the Optical Fiber Communication (OFC) conference, paper PD24
19. Essiambre R-J, Tkach RW (2012) Capacity trends and limits of optical communication networks. Proc IEEE 100:1035–1055
20. Zhou X et al (2011) 64-Tb/s, 8 b/s/Hz, PDM-36QAM transmission over 320 km using both pre- and post-transmission digital signal processing. J Lightwave Technol 29:571–577
21. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423
22. Essiambre R-J, Kramer G, Winzer PJ, Foschini GJ, Goebel B (2010) Capacity limits of optical fiber networks. J. Lightwave Technol. 28:662–701
23. Ryf R et al (2012) Mode-division multiplexing over 96 km of few-mode fiber using coherent 6 × 6 MIMO processing. J. Lightwave Technol. 30:521–531

24. Takara H et al (2012) 1.01-Pb/s crosstalk-managed transmission with 91.4-b/s/Hz aggregated spectral efficiency. In: Proceedings of the European conference on optical communications, paper Th3.C.1
25. Richardson DJ, Fini JM, Nelson LE (2012) Space-division multiplexing in optical fibres. Nat Photon 7:354–362
26. Li G, Bai N, Zhao N, Xia C (2014) Space-division multiplexing: the next frontier in optical communication. Adv Opt Commun 6:413–487
27. Fontaine NK et al (2015) 30×30 MIMO Transmission over 15 Spatial Modes. In: Proceedings of the Optical Fiber Communication (OFC) conference, post deadline paper Th5C.1

Optics in Remote Sensing

Thomas Walther and Edward S. Fry

T. Walther
Institute for Applied Physics, TU Darmstadt, Schlossgartenstr. 7, D-64289 Darmstadt, Germany

E.S. Fry (✉)
Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843-4242, USA
e-mail: fry@physics.tamu.edu

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_9

9.1 Introduction

The observation of light or more specifically the difference between day and night is the very first encounter of physics every human experiences at a very early stage in life. So not surprisingly, optics and the study of the properties of light is one of the oldest in the history of science. The concept of straight propagation of light in homogenous media dates back to the ancient Greeks. This is known as Hero's principle. Although we have learned a lot about light throughout the centuries and have discovered more subtle effects about how light propagates, Hero's principle is the most important feature when applying light to remote sensing. However, we will see that despite its relative early beginnings, modern remote sensing is intimately connected with modern tools of optics such as pulsed lasers, fast detectors, etc.

Optical remote sensing involves the use of light to observe distant objects and to obtain parameters or characteristics of those objects; it can be passive or active. Passive remote sensing would include seeing the light emitted by the object, e.g., a star, the sun, or even the headlights of an approaching car. Remote sensing in the latter case enables one to determine the location of the car and whether it may be on a collision course. Passive remote sensing would also include seeing an object, but via the light from some other source that is scattered/reflected by it; examples include seeing the moon and the planets via the light from the sun that is scattered by them, or just a boy standing in the street and illuminated by the afternoon sun. In either case the light scattered or emitted by the object can also be analyzed to determine some properties of the object; examples include the color of the object and how fast it is moving towards or away from you.

Active remote sensing involves sending out a beam of light and observing the light reflected or scattered by an object. A simple example would be the headlights of a car that illuminate a girl running across the street at midnight. Active remote sensing using a laser to illuminate the object is a technique named LIDAR, a name that arises from a linguistic blend of light and radar. In analogy with the term radar, one might say that it is an acronym for "LIght Detection And Ranging"; but there is no consensus and one can frequently find other definitions such as "Laser Imaging, Detection And Ranging," or "Light Intensity Distance And Ranging." In addition, there is no consensus that it should be an acronym with all capital letters; consequently, it is also variously referenced in publications as LiDAR, Lidar, or just lidar.

This contribution aims at outlining how light is used in remote sensing. First, a short historical overview of remote sensing will be given. This introduction is followed by some of the technical developments leading to modern remote sensing applications such as Lidar. Finally, we will explain some of the applications and advances in the field of Lidar. This article is intended to give a general overview of light in remote sensing while focusing on what is feasible today.

9.2 Historical Overview

The beginning of remote sensing is most likely the invention of triangulation, a technique to measure the distance to some location (possibly remote and inaccessible). Essentially, it involves selecting two positions whose separation is known and measuring the angle subtended at each position by the other position and by the remote location (via light coming from the remote location). Consequently, one has a triangle whose vertices are the two positions and the remote location; the length of one side is known and two angles have been measured, so simple geometry gives the distances to the remote location from each of the two positions. The ancient Greeks already knew that triangles can be used to estimate distances.

Thales had used such a technique in the sixth century BC in order to estimate the height of the Pyramids in Egypt. He simply compared the length of the Pyramid's shadow with his own shadow arriving at the ratio between the height of the Pyramid and his own height. While this knowledge was later somehow lost in Europe, it was noticed in ancient China around 200 AD that triangulation is essential to accurate cartography. The method of using triangles to measure distances was reintroduced to Europe by Arabic scholars around 1000 AD. During the sixteenth century triangulation was widely used in local cartography. The next big step forward was accomplished by Snell in the 1600s; he was the first to establish the required adjustments to the method in order to compensate for the Earth's curvature. Accordingly, local cartography was entirely based on triangulation and many countries established triangular networks over their entire landmass. Today, triangulation on a nationwide scale has been replaced by the global positioning system (GPS). In GPS, satellites continuously send timing information to the GPS receiver. As the exact positions of the satellites are known, the receiver can figure out the runtime differences for the signals to its location and thus determine the position of the receiver relative to that of the satellite. Thus, GPS is basically a modern variant of triangulation.

9.2.1 Speed of Light

In a sense, the measurement of the speed of light historically constituted a form of remote sensing. The ancient Greek philosopher Empedocles had already suspected that the speed of light is finite. However, at the time the theory that the eye actually sends out light was more common and so there was no real need for the notion of the speed of light. The Arabic scholar Alhazen published the first long treatise on optics and postulated that the eye actually only receives light. He then correctly argued that light has a finite speed and that in fact it travels slower in dense media. Since there was no clear idea how to experimentally prove that the speed of light was finite, over the centuries a lot of arguments were exchanged for either view. The first one to experimentally prove that light has a finite speed was Ole Rømer, a Danish astronomer, in 1676. He used an astronomical method in order to arrive at the conclusion that the speed of light was finite by observing the motion of the innermost moon of Jupiter, Io, on its orbit around Jupiter. Based on Rømer's observations, Huygens determined a value of the speed of light which differed from today's value by approximately one third. More refined techniques for measuring the speed of light were developed; these used a long optical path and rotating mirrors that would reflect the light back to the observer if the mirror had rotated to the right angular position by the time the light arrived at it. For a long time this had been the most accurate method until superseded by independent measurements of the wavelength and frequency of the light. The product of these quantities yields the speed of light.

Today, the speed of light has a defined value adopted in 1983. Since the invention of the atomic clock, time is the most accurately measured physical quantity. Consequently, the definition of length arrives at the same level of precision when the speed of light is fixed and distance is defined via the product of time and speed.

9.2.2 Fraunhofer and the Invention of Remote Sensing

Remote sensing in its modern meaning was unknowingly established by Fraunhofer. In 1801, Fraunhofer survived a collapse of the glass-making workshop in which he was working as an apprentice. The Prince Elector of Bavaria was so

elated to find at least one survivor that he donated money to the orphan Fraunhofer who was 14 years old at the time. Fraunhofer invested the money in books and learned in the following years to produce optical glass of previously unknown purity and quality. He further invented a spectroscope, essentially a predecessor of today's spectrographs, that can analyze light for its spectral content. In 1814, he observed almost 600 dark lines in the white solar spectrum—far more than Wollaston had observed a couple of years earlier and independently from Fraunhofer. Later Bunsen and Kirchhoff realized that these lines actually were absorption lines of gases in the solar atmosphere. Kirchhoff recognized that a particular set of lines was unique to one special element just like a fingerprint is for a human being. This essentially established spectroscopy as a science and constitutes the first example of remote sensing of the gases in the solar atmosphere. In honor of Fraunhofer, these absorption lines are called Fraunhofer lines. This very specific absorption of any gas (atoms or molecules) will prove very useful when we discuss active remote sensing schemes below.

9.2.3 Passive Remote Sensing

The invention of photography in combination with the ability to fly in unmanned or manned vehicles led to the birth of a broad range of passive remote sensing. The first example is aerial photography. Needless to say, this was very important for military purposes; but it was soon realized that other very valuable information could be extracted from the pictures. This is true in particular for the much higher quality images available from high-flying planes and nowadays satellites. The very sensitive detection devices and optics on board modern satellites allow a variety of remote sensing applications. But, since this article is focussed on optics in remote sensing, the discussion will omit techniques applying acoustic or radar based methods. The most common objectives in the passive satellite based remote sensing programs include meteorological observations for weather forecasting, predictions of hurricane movements, sea surface temperature distributions, military applications and reconnaissance, topographic mapping (stereo photography), or applications in mineralogy, biology, and archaeology.

One of the more sophisticated techniques is hyperspectral imaging in which a variety of images in different spectral regions are obtained; these give a complete picture of the reflective behavior of the Earth's surface. Thus, information such as plant coverage, snow coverage in mountains or the Arctic, and ground-level humidity can be extracted from the data. Such information is especially useful in, for example, analyzing the environmental impact of human activities, or in monitoring other environmental changes.

9.3 The Development of the Laser for Active Remote Sensing

The big leap from passive to active remote sensing required an appropriate light source capable of sending collimated light beams over large distances. In 1960 such a light source, the laser, was operated for the first time; it was invented by T. Maiman after C. Townes and A. Shalow had shown in 1958 that it was theoretically feasible. The acronym laser stands for light amplification by stimulated emission of radiation. In 1918 Albert Einstein had investigated a different phenomenon associated with the interaction of light with matter. Light could be absorbed by an atom, i.e., if a photon, whose energy corresponded to the energy difference between the ground and an excited state of the atom, were to

interact with the atom, it could be annihilated and the atom would then be elevated to the excited state of higher energy. The reverse process of an atom in the excited state falling back down to the ground state would then lead to the emission of a photon. Essentially, there are two kinds of emission processes: spontaneous emission occurring by chance without any cause, and stimulated emission occurring when a photon of the correct energy difference interacts with an excited atom and stimulates it to emit. In the latter case, the final result is an atom in the ground state and two photons—the original one and the stimulated one; furthermore, the stimulated photon is an exact copy of the original photon. In essence, this is an amplification process. In the laser this process leads to a large increase in the number of photons; specifically, since the photons are reflected back and forth many times through the medium, many photons are generated. Since the probability an incident photon will be absorbed by a ground state atom is identical to the probability an incident photon will stimulate an excited atom to emit a photon, this laser process works when there are a larger number of atoms in the excited state than in the ground state. This condition is referred to as a population inversion. The many reflections are made possible by the fact that the medium is between two mirrors forming a resonator. One of the mirrors is slightly less reflective than the other so that a small amount of light leaks through it, leaves the laser resonator, and can be used for applications. There are three important properties of this light: First, since the resonator defines an axis of symmetry, the light is highly directed along this very axis; second, since the radiation is closely linked to an atomic transition, the light consists of a single frequency corresponding to that atomic transition frequency; and third, since the photons are generated in a stimulated emission process, the light is coherent. All three properties make up the very unique features of laser light and essentially make it ideal for our purpose of remote sensing. However, just the invention of the laser was not yet enough to fulfill all our needs.

The next necessary step was the invention of the so-called Q-switched laser, which followed in 1962 very soon after the invention of the laser. While a laser in itself can produce very powerful light, the amount of light emanating from it was not enough to perform remote sensing in the atmosphere. The "Q" in Q-switch stands for quality. When the medium of a laser placed inside a cavity is pumped by an energy source achieving population inversion, light is amplified as soon as the gain of the stimulated emission exceeds the losses of the cavity. Consequently, the overall energy in the system is reduced by the stimulated emission (lasing) leaving the medium in the cavity. However, the medium could potentially store much more energy if lasing did not occur.

So, what if at first the cavity was of low quality? Light is not reflected back and forth and no amplification takes place. In this case, almost all of the energy in the pump source is stored in the medium. Once this has happened, the cavity can be suddenly switched into a state of high quality. Now, all of the energy stored in the medium is deposited into a giant pulse with very high energy and a few nanoseconds duration, i.e., some billionths of a second. The sudden switch of the cavity's quality is performed by the Q-switch. In general, this is an electro-optic modulator that manipulates the polarization of the light by applying a short voltage pulse to it.

Inevitably, a pulsed laser has a larger linewidth than a continuous wave laser. The minimum bandwidth is given by the so-called Fourier transform limit. In general, however, a pulsed laser has an even larger bandwidth than this limit. But, since the transitions in typical trace gases are fairly well separated, this linewidth increase does not change the selectivity in detection of trace gases. As stated above, the typical pulse durations of Q-switched lasers are in the nanosecond regime. This duration actually determines the typical spatial resolution one can achieve via time-of-flight measurements.

The first laser in 1960 was based on a solid state material known as ruby. The first Q-switched laser was based on the very same material. Ruby as a laser material, however, has a serious drawback. It is only capable of generating laser light around a very well-defined frequency. In fact, most of the early lasers had this limitation, albeit at different wavelengths. Unfortunately, these available laser frequencies were not resonant with any of the transitions of the interesting gases for remote sensing. Thus, another step was necessary in the development of the laser as a valuable tool in remote sensing. The next advancement was the invention of the tunable laser, i.e., a laser whose output frequency could be changed so as to match the transitions of gases of interest. The first tunable laser was the dye laser invented in 1966 independently by Sorokin and by Schäfer. In a dye laser the laser medium consists of an organic dye dissolved in an organic liquid such as methanol and ethanol. In solution, the individual resonances of the dye broaden to a wide band that leads to a broad emission spectrum. By placing additional optical components in the cavity, the laser can be restricted to work at a particular frequency within this band. And, by slightly adjusting the alignment of the components, the output wavelength can be tuned over the entire emission band of the dye. Now, the output of the laser could be tuned to the exact transition needed for a particular gas to be detected. In the following years many dyes were found suitable for such lasers; they cover a broad range of different wavelengths from the IR range, through the visible spectrum, and into the UV regime. And, the Ti:Sapphire laser now provides a solid state alternative that is capable of generating any wavelength in the near IR range.

As we will see later, even shorter pulses can be advantageous for remote sensing. The reason is twofold. First, very short pulses have very high peak intensities that can be directly applied to non-linear detection schemes that are now both available and increasingly important for applications in remote sensing. Second, since pulsed lasers have a larger bandwidth, there is the possibility of detecting several different species at once. Unfortunately, Q-switching only provides pulses in the nanosecond regime, this is essentially due to a combination of the typical size of the laser cavities and the gain of the media involved. However, a laser resonator does not lase on any wavelength, but only on the so-called longitudinal modes. This is comparable to the oscillations of a mechanical string fixed at both ends. The length of the string is an integer multiple of half the wavelength of the possible oscillations; the same is true of the laser oscillations. Thus, in general, the spectrum of a pulsed laser consists of a superposition of these longitudinal modes. There are techniques to synchronize these different modes. When these modes all oscillate with the same phase, something remarkable happens: the radiation in these modes interferes constructively to form giant pulses. The more longitudinal modes are involved the shorter the duration of these pulses. Moreover, the giant pulses oscillate back and forth in the cavity and exit at the corresponding high repetition rate.

9.4 LIDAR

Lidar for active optical remote sensing has a broad range of important applications. Some specific applications to be discussed in the following include: (Sect. 9.4.1) the precision measurement of distances over ranges varying from a few meters to thousands of kilometers, (Sect. 9.4.2) measuring the speed of an object at a distance point, (Sect. 9.4.3) measuring sound speed as a function of depth in the ocean, (Sect. 9.4.4) measuring temperature as a function of depth in the ocean, (Sect. 9.4.5) detecting and identifying underwater objects (fish, mines, etc.), (Sect. 9.4.6) detecting trace impurities in the atmosphere, (Sect. 9.4.7) a quite recent development, a femtosecond-Lidar application for influencing

weather phenomena, and (Sect. 9.4.8) stand-off super-radiant spectroscopy. A few other important applications that require at least a brief mention include steering a laser beam by inducing refractive index gradients, measuring depth profiles of particulate backscattering in the ocean; obtaining broad areal depictions of the sound speed environment; unraveling sonar signals in regions of varying sound speed; collecting important input for weather forecasting and climate change studies; understanding the behavior of biological populations in the ocean as well as the interactions between oceanic physical and biological structures; and studying/mapping the atmosphere for gases, aerosols, clouds, and temperature.

9.4.1 The Precision Measurement of Distances

The distance d to an object can be measured by sending a short laser pulse to the object and then measuring the time t it takes for the reflected pulse to return. Since the pulse travels a round-trip distance $2d$ in time t, the distance d is given by

$$d = \frac{ct}{2}, \tag{9.1}$$

where c is the speed of light. In practice, one sends a series of pulses, measures the travel time t for each pulse with sub-nanosecond accuracy, and averages the results to then obtain d from Eq. (9.1).

This approach to measuring distances has developed a wide range of applications. It is, for example, very useful in construction and real estate; it essentially replaces the tape measure. Rather than two people with a tape measure to determine the size of a large room, one person with a laser rangefinder can quickly measure it to very high accuracy. Typical of rangefinders for this use is the Bosch Model GLR825 which can measure distances from 50 mm to 250 m with an accuracy of ±1 mm [1]; it presently sells for about $400. Use of a tape measure to determine the width of a 10 m wide room to this kind of accuracy (i.e., one part in 10,000) would be highly problematic.

At the other extreme of LIDAR distance measurements, is the determination of the distance from the earth to the moon (the mean value is 385,000 km). The astronauts placed retroreflectors (corner cubes that reflect incident light by 180°) on the moon during the first lunar landing in 1969; there have been several other retroreflectors placed on the moon since then. The time t for a light pulse to travel from the earth to those retroreflectors and back is then measured and the distance from the surface of the earth to the surface of the moon is determined using Eq. (9.1). Such measurements are quite challenging. As discussed by Dickey et al. [2], the area illuminated on the lunar surface due to the divergence of the transmitted laser beam is ~7 km diameter and the retroreflector only intercepts $\sim 10^{-9}$ of that area. Due to the diffraction/divergence of the retroreflected beam, a 1 m diameter telescope on the earth will only collect $\sim 2 \times 10^{-9}$ of it. These two factors lead to a signal collection that is only 2×10^{-18} of the initial laser beam intensity I_0. Other factors such as detector quantum efficiency and mirror reflectivities less than one lead to an overall observed signal of $\sim 10^{-21}\ I_0$. So, if the laser sends out 10^{19} photons in each pulse, the average observed signal would only be one retroreflected photon for every 100 pulses. Nevertheless, the earth–moon distance can be measured with such high accuracy (~1 cm) that the earth–moon system becomes a "laboratory for a broad range of investigations, including astronomy, lunar science, gravitational physics, geodesy, and geodynamics" [2].

Some additional examples in this wide range of applications for laser rangefinder measurements include:

1. Autonomous vehicles—they require knowledge of the distance to nearby objects [3, 4] in order to avoid collisions.
2. Military applications—if the distance to a target is so great that gravitational effects on the projectile must be taken into account when aiming at the target [5, 6], then knowledge of the distance is critical. The MARK VII military rangefinder is an example of a device that meets this need; it operates at ranges of up to 20 km with an accuracy of ±3 m [7].
3. Sports—in golf, the use of laser ranging can significantly improve hitting accuracy by providing an accurate distance to the target [8].
4. Archaeology—high resolution depth measurements and contour mapping at a site can reveal archaeological features that are otherwise hidden [9].
5. Sub-surface topographical measurements—similar to the archaeology application, the topology of reef substrates (that impact the biology of reef organisms) can be obtained using NASA's Experimental Advanced Airborne Research Lidar (EAARL) [10, 11].

9

9.4.2 Measuring the Speed of an Object at a Distance Point

If a laser beam is reflected back on itself from an object that is moving in either direction along the same line as the laser beam, then the frequency of the reflected light will be Doppler shifted. By measuring the frequency shift of the reflected light, one can obtain the component of the object's velocity parallel to the laser beam and also determine whether it is approaching or going away. If the object is moving perpendicular to the laser beam, there is no frequency shift in the retroreflected light; but, this component of the velocity could be measured by measuring both the distance and the changing angular direction to the object. The bottom line is that the frequency shift only gives the component of the object's velocity along the direction of the laser beam.

In practice, the system works just as with the rangefinder; a laser pulse is directed to a target and the reflected light is collected. The travel time can still be used to determine the distance to the target; but, in this case the reflected light is also analyzed in frequency to determine the Doppler shift.

To quantify the speed as a function of the frequency shift, recall that the observed frequency due to the motion of a source of electromagnetic waves relative to the observer is given by [12]

$$f = f_0 \sqrt{\frac{c+v}{c-v}}, \tag{9.2}$$

where f is the frequency measured by the observer, f_0 is the frequency in the rest frame of the source, c is the speed of light in the surrounding medium, and v is the relative speed of the source and observer—it is positive (negative) when the source is approaching (moving away from) the observer. Note, v is just the relative speed of source and observer; it does not matter whether the source, observer, or even both are moving.

Now consider the LIDAR application to determine the speed of a target. The laser is the source with frequency f_0, the target is the observer that sees a frequency f_T given by Eq. (9.2) (f on the left-hand side is replaced by f_T). The target is now the

source and reflects source light of frequency f_T back towards the laser; the laser position is now the observer and observes a return frequency f_R given by

$$f_R = f_T\sqrt{\frac{c+v}{c-v}} = f_0\sqrt{\frac{c+v}{c-v}}\sqrt{\frac{c+v}{c-v}} = f_0\left(\frac{c+v}{c-v}\right) \tag{9.3}$$

If $v \ll c$, this can be approximated as

$$f_R \approx f_0\left(1+\frac{2v}{c}\right), \tag{9.4}$$

and the frequency shift due to the Doppler effect is

$$f_R - f_0 \approx \frac{2v}{c}f_0 = \frac{2v}{\lambda_0}. \tag{9.5}$$

Suppose a target is moving with a speed $v = 5$ m/s (~11 miles/hr) and the laser has a wavelength $\lambda_0 = 500$ nm (~blue). Then, the Doppler frequency shift from Eq. (9.5) is

$$f_R - f_0 = 20 \text{ MHz}. \tag{9.6}$$

Measuring such a frequency shift with a good Fabry–Perot is straightforward; a more difficult problem is the laser bandwidth. For a Gaussian laser pulse, the Fourier transform limited time-bandwidth product is

$$\tau\,\Delta f \geq \frac{2\ln 2}{\pi} \approx 0.44, \tag{9.7}$$

where Δf is the frequency bandwidth and τ is the temporal width (both full width half maximum). Thus, in order to resolve the Doppler shift in Eq. (9.6), the temporal width of the laser pulse must be greater than $\tau = 22 \times 10^{-9}$ s. Of course, for targets moving at higher speeds, this becomes less of a limitation. If the target is moving at 50 m/s, the Doppler frequency shift is 200 MHz and the laser pulse would only be required to have a temporal width greater than 2.2 ns.

9.4.3 Measuring Sound Speed as a Function of Depth in the Ocean

A light pulse propagating in the ocean is scattered by density fluctuations that propagate with the speed of sound. This scattering process is called Brillouin scattering and the remote sensing application, called Brillouin Lidar, has been studied [13–15]. As before, the depth is determined from the travel time of the light pulse; the speed (of sound) is measured by observing the Brillouin frequency shift f_B in the backscattered light. The basic physics of the Brillouin scattering problem is shown in Figs. 9.1 and 9.2.

Figure 9.1 shows the sound wave fronts for sound waves of wavelength λ_S and wave vector $\vec{k}_S$. A light beam of wave vector $\vec{k}$ is incident at an angle $\theta/2$ with respect to the sound wavefronts. The scattered light has a wave vector $\vec{k}'$; but since $\left|\vec{k}'\right| \approx \left|\vec{k}\right|$ and the angle of incidence on the sound wavefronts is equal to the angle of reflection, the scattered light must be at angle θ with respect to the incident light beam. From the sub-diagrams in Fig. 9.1, it is clear from momentum conservation that the magnitudes of the wave vectors must satisfy

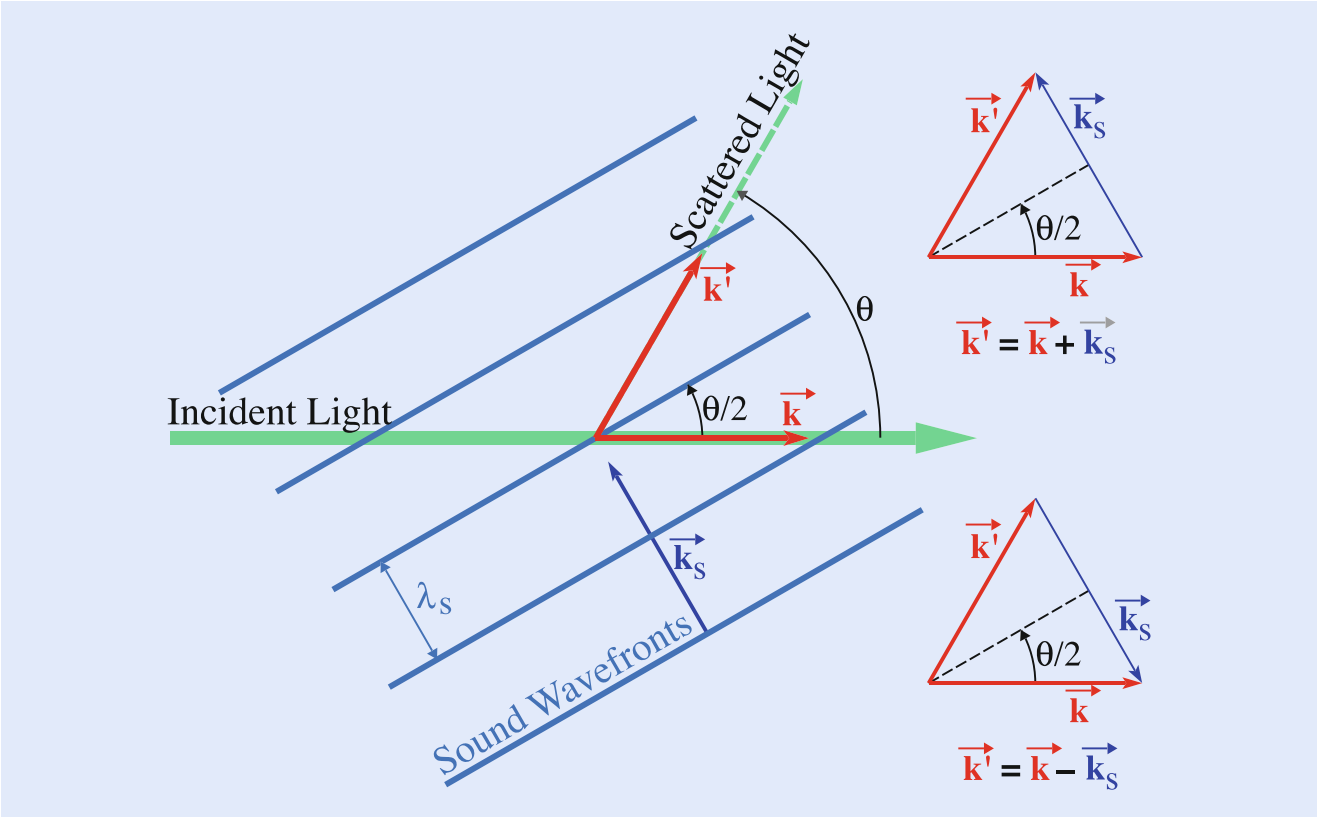


Fig. 9.1 Vector diagram for momentum conservation in Brillouin scattering

$$k_S = 2k \sin\frac{\theta}{2}. \tag{9.8}$$

Figure 9.2 shows a sound wavefront that moves a distance $v_S\Delta t$ in a time Δt. Point a on the initial sound wavefront becomes point b after the wavefront has moved the distance $v_S\Delta t$. Since the reflecting surface is a plane wavefront, the object and image distances must be the same. Thus, the distance of the source plane from point a equals the distance of the initial image of the source plane from point a. Similarly, the distance of the source plane from point b equals the distance of the image of the source plane from point b after the time Δt. From the inset in Fig. 9.2 it is clear the distance $V\Delta t$ that the image plane moves in time Δt must be given by

$$V\Delta t = 2\left(v_S\Delta t \sin\frac{\theta}{2}\right). \tag{9.9}$$

The relative speed V between the fixed observation plane and the moving image of the light source is therefore

$$V = 2v_S \sin\frac{\theta}{2}. \tag{9.10}$$

From Eq. (9.2), the observed frequency due to a source moving with a speed of magnitude V is

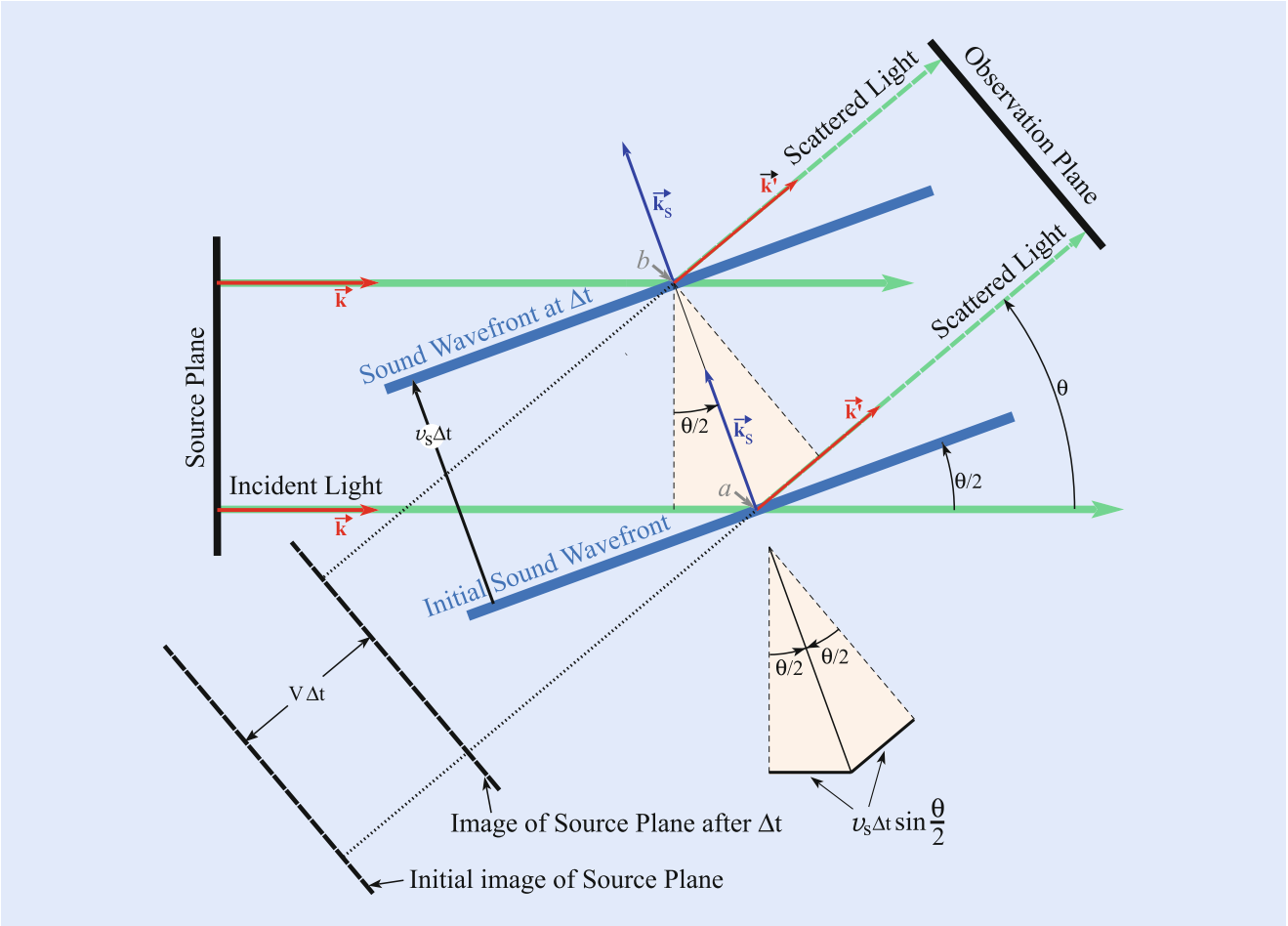


Fig. 9.2 Diagram for determining the Doppler shift

$$f_{\text{obs}} = f_0 \sqrt{\frac{c \pm V}{c \mp V}} \approx f_0 \left(1 \pm \frac{V}{c}\right), \tag{9.11}$$

where the approximation in the final step is that $V \ll c$. The upper (lower) sign corresponds to the image of the source plane moving towards (away from) the observation plane. The upper sign would be for $\vec{k}_S$, as shown in Fig. 9.2; the lower sign would be for $\vec{k}_S$, as shown in the second of the sub-diagrams of Fig. 9.1. The Brillouin frequency shift f_B is then

$$f_B = f_{\text{obs}} - f_0 = \pm \frac{V}{c} f_0 = \pm \frac{nV}{\lambda_0}, \tag{9.12}$$

where n is the index of refraction of the medium and λ_0 is the vacuum wavelength of the incident light. Inserting V from Eq. (9.10) gives

$$f_B = \pm v_S \, 2 \frac{n}{\lambda_0} \sin \frac{\theta}{2}. \tag{9.13}$$

Note that for backscattering ($\theta = 180°$) with $n = 1$, Eq. (9.13) is identical to the Lidar frequency shift of an object moving at speed v_S, Eq. (9.5). Recall that the

wave vector has a magnitude given by $k = 2\pi n/\lambda_0$ for the incident light and $k_S = 2\pi/\lambda_S$ for the sound waves. Insert these in Eq. (9.8),

$$\frac{1}{\lambda_S} = 2\frac{n}{\lambda_0}\sin\frac{\theta}{2}, \tag{9.14}$$

and combine this result with Eq. (9.13) to obtain the interesting result,

$$f_B = \pm\frac{v_S}{\lambda_S} \equiv f_S, \tag{9.15}$$

that these two frequencies are identical, i.e., the Brillouin frequency shift of the incident light is identical to the frequency of the sound wave producing it.

The speed of sound waves is obtained by simply measuring the Brillouin shift f_B at some angle θ (generally backscattering at 180°) and using Eq. (9.13). For $\theta = 180°$, the sound speed is given by

$$v_S(S,T) = \frac{\lambda_0}{2}\frac{|f_B(S,T,\lambda_0)|}{n(S,T,\lambda_0)}$$

9

where the dependencies on salinity S, temperature T, and wavelength λ_0 have been explicitly shown.

For a typical value of the Brillouin shift, use Eq. (9.13) and take $\theta = 180°$, $\lambda = 530$ nm, $n = 1.33$, and $v_S = 1500$ m/s; the result is

$$f_B = \pm 7.5 \text{ GHz}. \tag{9.17}$$

In the oceans the Brillouin shifts are typically in the range of 7–8 GHz. In a Brillouin LIDAR, the sound speed actually being measured is the speed of high frequency (~7.5 GHz) sound waves.

As an example of a Brillouin spectrum, Fig. 9.3 shows experimental data for Brillouin backscattering by water using the second harmonic of a pulsed Nd:YAG laser [13]. The central peak (at relative frequency 0) is due almost entirely to elastic scattering by suspended particulates. The two peaks offset by ~7.5 GHz are the two Brillouin shifted peaks; they are due only to the water—suspended particulates do not contribute to them. The ratio of the intensity in the central peak to the total intensity in the Brillouin peaks is called the Landau–Placzek ratio [16, 17]. An interesting property of high purity water is that this ratio is so small;

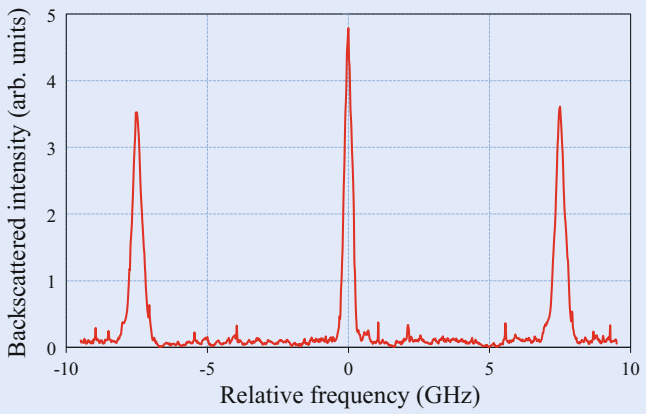


Fig. 9.3 Experimental data of Brillouin backscattering by water

experimentally and theoretically, it is a minimum and approximately 0–0.04 % at 4 °C, it increases with temperature and is approximately 2 % at 30 °C [17, 18].

A practical implementation of this Brillouin Lidar concept requires a receiver that collects light over an appreciable solid angle, that provides a high frequency resolution of the ~7 GHz frequency shifts, and that provides the measurements with ≈10 ns resolution over a time scale of several hundred nanoseconds (10 ns gives a depth resolution in water of ~1 m).

The first approach to achieve the required resolution might be a Fabry–Perot, but unfortunately the angular acceptance of a suitable Fabry–Perot is too small. For a Fabry–Perot with mirror separation d, the optical path difference between successive rays is $\delta = 2nd\cos\theta$, where n is the index of refraction of the medium between the plates and θ is the angle of incidence on the mirror surface inside the Fabry–Perot. At the transmission peaks the path difference is

$$\delta = 2nd\cos\theta = N\lambda_0, \tag{9.18}$$

where λ_0 is the vacuum wavelength and N is an integer which has its maximum value when $\theta = 0$.

If the free spectral range is $\Delta\lambda$ and the fullwidth at half maximum of the transmission peaks is γ, then the finesse F is defined as $F = \Delta\lambda/\gamma$. As the angle θ increases from the angle of a transmission peak, the path difference decreases and the transmission goes to zero, but rises again at the next transmission peak when the path difference has decreased by λ_0 and Eq. (9.18) is satisfied using $N-1$. If θ is increased from $\theta = 0$, then the maximum angle θ_m at which the transmission moves off the peak occurs when the path difference has decreased by a fraction of λ_0 given by $\gamma/\Delta\lambda$,

$$2nd\cos\theta_m = N\lambda_0 - \frac{\gamma}{\Delta\lambda}\lambda_0 = N\lambda_0 - \frac{\lambda_0}{\mathcal{F}} = 2nd - \frac{\lambda_0}{\mathcal{F}}. \tag{9.19}$$

Solving for θ_m gives

$$1 - \cos\theta_m = \frac{\lambda_0}{2nd\mathcal{F}}, \tag{9.20}$$

and expanding $\cos\theta_m$ for small θ_m gives

$$\theta_m \approx \sqrt{\frac{\lambda_0}{nd\mathcal{F}}}. \tag{9.21}$$

For typical values of $d = 1$ cm, $F = 40$, $n = 1$, and $\lambda_0 = 532$ nm, the angle is $\theta_m \sim 0.07°$; for reference, the angles for the transmission peaks are 0°, 0.42°, 0.59°, 0.72°, etc. The problem with $\theta_m \sim 0.07°$ can be understood by considering a practical situation in which the Brillouin Lidar receiver mirror has a diameter of 80 cm. The maximum divergence of Lidar signals collected by this mirror from a point source at a distance of 150 m would be 0.31°. This is already too large compared to 0.07°, but the problem is even worse. Specifically, a telescope would be required to reduce the diameter of the Lidar return from 80 cm to the few cm diameter of a practical Fabry–Perot; leading to a further significant increase in the beam divergence. The problem has been examined from several aspects by Hickman et al. [14].

The edge technique, a more recent approach, provided the capability to collect the Brillouin scattered light over an appreciable solid angle, to obtain the Brillouin frequency shift as a function of time with ≈10 ns resolution (i.e., 1 m depth resolution), and to do this over a time interval of hundreds of nanoseconds.

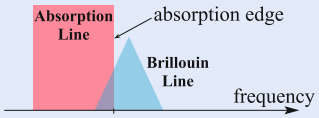


Fig. 9.4 The edge technique concept using an atomic/molecular absorption line

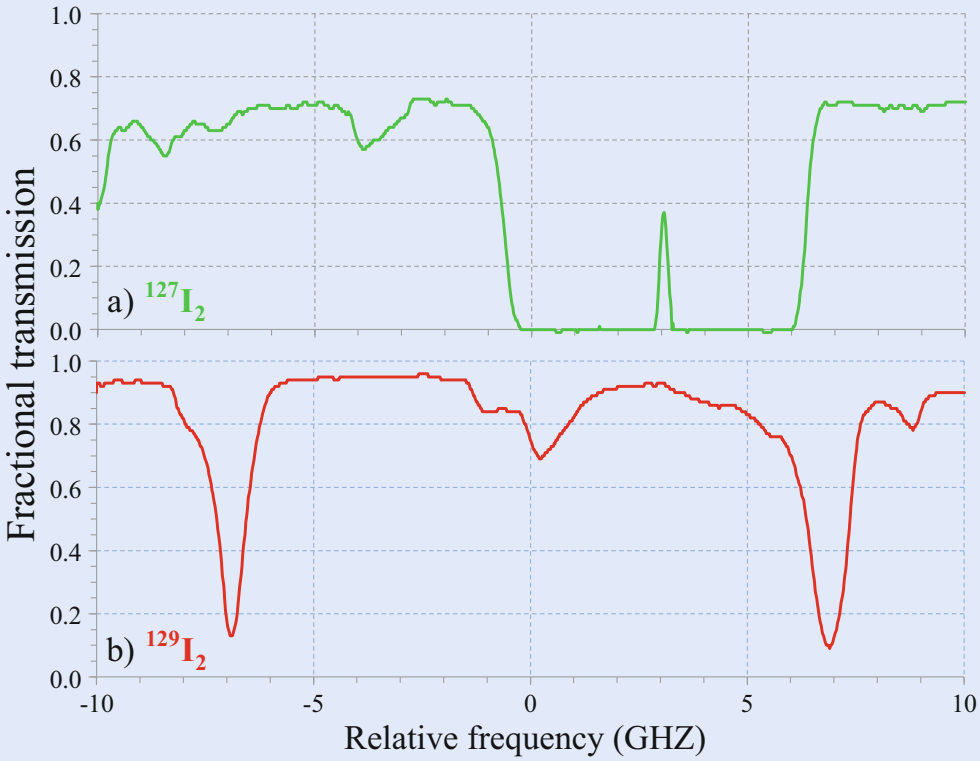


Fig. 9.5 Absorption spectra for the molecules (**a**) $^{127}I_2$ and (**b**) $^{129}I_2$; the zero on the relative frequency axis corresponds to an Nd:YAG laser wavelength of 532.38 nm

The approach is to use an atomic/molecular absorption line and to choose a laser frequency so that a Brillouin shifted component will lie on the edge of the absorption line. Figure 9.4 illustrates the concept with a fixed frequency rectangular absorption line and a triangular Brillouin line that partially overlaps the absorption line. The Brillouin Lidar return passes through a cell containing the absorbing gas that partially absorbs the Brillouin scattered light. The absorption decreases as the Brillouin shift increases, and vice versa.

Of course, the practical implementation is more complicated because there are two Brillouin lines that must lie on the edges of absorption lines; and there is also the central unshifted line due to scattering by particulates in the water that must be removed. Molecular absorption spectra were examined and a set of absorption lines that meet the requirements were found in $^{127}I_2$ and $^{129}I_2$; their absorption spectra are shown in Fig. 9.5.

The $^{127}I_2$ has strong absorption at the relative frequency zero (corresponding to the second harmonic of a Nd:YAG laser at 532.38 nm); it is used to remove the central peak. The outside edges of the two $^{129}I_2$ absorption lines are at a relative frequency of approximately ±7.5 GHz which would be the frequency of the

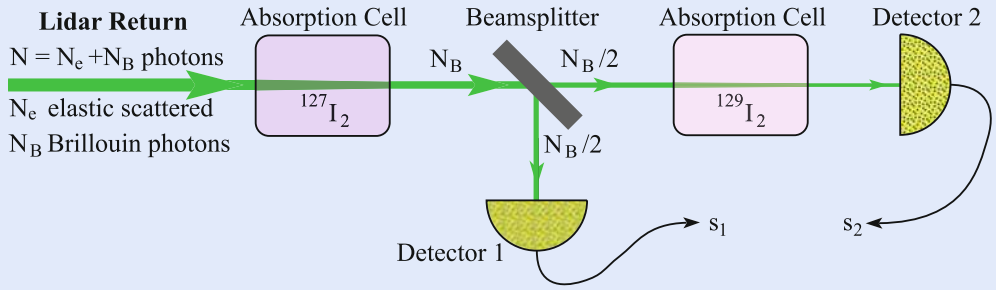


Fig. 9.6 Brillouin Lidar receiver using in $^{127}I_2$ and $^{129}I_2$ absorption cells

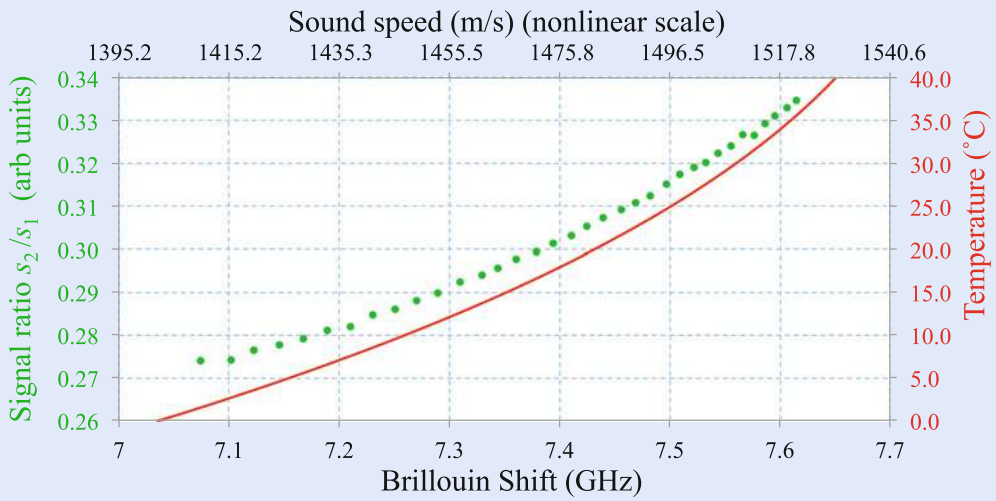


Fig. 9.7 Signal ratio s_2/s_1 (*dots*) and temperature (*smooth curve*) as a function of the Brillouin shift for pure water ($S = 0$); note that the ratio s_2/s_1 is on an arbitrary scale

Brillouin scattered peaks for typical conditions of temperature and salinity in the oceans and other natural waters.

A Brillouin Lidar receiver that implements the edge concept using these molecular absorption lines is shown in Fig. 9.6. The Lidar return first passes through a $^{127}I_2$ absorption cell that removes the central component and leaves only light that has been Brillouin scattered. Half of this light is then sent to a detector to obtain a reference signal s_1 for the total Brillouin scattered light. The other half passes through a $^{129}I_2$ absorption cell that absorbs a fraction of the Brillouin scattered light based on the extent to which the Brillouin lines overlap the $^{129}I_2$ absorption lines; it provides a signal s_2. After calibration, the ratio s_2/s_1 provides the Brillouin shift.

The efficacy of the system shown in Fig. 9.6 is demonstrated by its application to obtain the data in Fig. 9.7 which shows the measured ratio of s_2/s_1 (*dots*) as a function of the Brillouin shift at the laser wavelength, $\lambda_0 = 532.57$ nm, and salinity $S = 0$. The sound speed at each Brillouin shift marker is shown at the top of the plot; this scale is not quite linear because of the temperature dependence of the index of refraction. In practice, the temperature was varied and used to calculate the resulting Brillouin shift and sound speed at each data point [19]. The relation between temperature and Brillouin shift is also shown in Fig. 9.7 (smooth curve).

The use of the edges of molecular absorption lines was the first approach to the edge filter concept. But an especially promising approach is an excited state Faraday anomalous dispersion optical filter (ESFADOF) [20–22]. Basically, an absorption cell containing an atomic vapor is placed between crossed polarizers and in a strong magnetic field. The high anomalous dispersion in the vicinity of the absorption line rotates the polarization so some fraction of the light (based on its frequency overlap with the edge of the absorption line) passes through the second of the crossed polarizers. The ESFADOF can provide sharp absorption edges at the desired frequencies by adjusting the strength of the magnetic field. There has already been considerable progress made with this concept using the 543.3 nm transition ($5S_{1/2} \rightarrow 8D_{5/2}$) between excited states in Rubidium [23].

Stimulated Brillouin scattering (SBS) has also been used to measure Brillouin shifts [24, 25]. This approach gives a larger signal and since the source of SBS is small, the scattered light can be collimated to within the acceptance angle of a Fabry–Perot. The latter provides the spectral distribution of the SBS and hence a measurement of the Brillouin linewidth; this could be especially useful because at temperatures below 10 °C the Brillouin linewidth has a strong dependence on temperature [26]. SBS does not produce the anti-Stokes line, but its serious drawback is that it makes a measurement at one specific depth where the incident beam is focused to sufficient intensity to produce SBS. Consequently, depth profiling of sound speed requires refocusing of both the laser transmission and the receiver for each depth. In another application, SBS has recently been used to measure the bulk viscosity of water [27]; a measurement that had initially been demonstrated with spontaneous Brillouin scattering [28].

In summary, the very real possibility of sound speed profiles in the top 100 m of the oceans promise vital information to oceanographers, biologists, environmentalists, the military, and others. Aircraft could be used to rapidly obtain sound speed profiles from which the sound speed structure could be determined for large areas of the upper 100 m of the ocean. Knowing this structure would have many important applications; it could, for example, provide information about acoustic surface ducts and their temporal changes.

9.4.4 Measuring Temperature as a Function of Depth in the Ocean

Since over 70 % of the earth's surface is ocean, its temperature distribution on the surface and throughout the top mixed layers plays a major role in our weather and in forecasting our weather. Those temperature distributions in the upper-ocean mixed layers are also of major importance to understanding the physical and biological behaviors of the ocean. Fortunately, experimental depth profile measurements of the Brillouin shift can be converted to depth profiles of the temperature. Actually, this is already clear from Fig. 9.7—a given ratio s_2/s_1 (*black dot*) corresponds to a unique Brillouin shift (*bottom* axis), a unique sound speed (*top* axis), and a unique temperature (the temperature given by the *smooth curve* for that unique Brillouin shift and sound speed). It is also clear from Eq. (9.13), which in analogy with Eq. (9.16), can be written as

$$f_B(S, T, \lambda_0) = \pm 2 \frac{n(S, T, \lambda_0) v_S(S, T)}{\lambda_0}. \tag{9.22}$$

Since λ_0 is known, if f_B is measured, then Eq. (9.22) is a relationship between salinity S and temperature T. If salinity is known to within 0.1 %, a theoretical analysis shows that an uncertainty of 1 MHz in the Brillouin shift measurement corresponds to a temperature uncertainty of approximately 0.07 °C [19]. If the

value of salinity is based on an historical compilation of data [29], it can be expected to be accurate to within a standard deviation of about 1 %; with a 4 MHz uncertainty in the Brillouin frequency shift, the temperature uncertainty would be approximately 0.5 °C [19] (a smaller uncertainty in the Brillouin frequency shift would not have much effect on the temperature uncertainty).

For simplicity, the above discussion neglected pressure effects, but they are fairly easy to include when needed. In principle, the pressure P affects both the index of refraction $n(S,T,\lambda_0,P)$ and the sound speed $v_S(S,T,P)$; but the effects of pressure on the index of refraction are essentially negligible—at 0.0 °C the index increases by less than 0.015 % for a pressure increase of 10 atm [30]. However, pressure can have significant effects on sound speed; it has been well documented and an empirical equation for the dependence is available [31]. This equation must be used in the above calculations for $v_S(S,T,P)$ when the pressure differs from 1.0 atm.

9.4.5 Detecting and Identifying Underwater Objects (Fish, Mines, etc.)

There is an interesting and dramatically simpler version of the Brillouin Lidar that provides the capability of detecting and possibly identifying underwater objects [32, 33]. The point is that the Brillouin Lidar return consists of a central peak due to elastically scattered light and two Brillouin shifted side peaks. Those side peaks only occur if light is being scattered by water at the depth of observation. If there is anything at that depth other than water, the side peaks will not appear.

To detect and identify an underwater object, the laser is tuned to, for example, a strong I_2 absorption line that is chosen based on the condition that I_2 has negligible absorption in the spectral region between 7 and 8 GHz on either side of that absorption line. The Lidar return is then passed through a cell containing I_2 vapor where all elastically scattered light is removed from the Lidar return; only Brillouin scattered light remains. If there is something other than water at the observation depth, there is no light in the Lidar return. It does not matter what that something is, if it is not water, there is no light.

In practice, the Lidar return would pass through range gate optics to define the observation depth in the water, then through the I_2 cell to absorb all elastically scattered light, and then into a camera. The camera would produce a relatively uniform picture if nothing is in the water. But if there is, for example, a fish at the observation depth, the picture would be uniform except for a black region corresponding to the outline of the fish. For a single laser shot, the problem would be for the camera (or array detector) to collect enough light to make a good picture.

9.4.6 Trace Gas Detection

As we have seen earlier, the principle of Lidar can be used to accurately measure distances. Combining this capability with the very specific absorption features of gases provides a method to determine trace gas concentrations in the atmosphere. The idea is simple: tune the pulsed laser to the resonance frequency of the trace gas of interest and fire a short pulse of laser light into the atmosphere. The laser will be scattered back towards the sender. The time of flight of the backwards scattered light will yield the distance at which the pulse was scattered. In this way, a short pulse of nanosecond duration will be broadened to microseconds. But the light returning in a particular nanosecond slice within this broader pulse returns from a

well-defined distance. A requirement for this to be true is that the scattering cross-section be small enough such that there are no multiple-scattering events. If there were any, ambiguities in the return time would be introduced. When a certain amount of the trace gas is found in the path of the laser radiation, the light is absorbed more efficiently, indicating that indeed there is a certain amount of the trace gas in the path. In the returning pulse the trace gas would manifest itself as a reduction in the signal. As an example, consider the trace gas to be present only at a height of 10–15 km then obviously light coming back very early, i.e., from heights less than 10 km, no absorption would be present. The same is true for the return signal from heights above 15 km, but of course this light also has to propagate twice through the absorbing layer. Thus, a reduced signal would be observed in the return signal starting at times corresponding to a distance of 10–15 km. This reduction is stronger than the regular signal decay due to a longer path through the atmosphere. However, there is a serious problem. What if clouds are present at a height of 5–10 km? Clearly, the clouds lead to increased levels of scattering. Less light returns to the observer, and the observer not knowing about the clouds would attribute the absence of the signal to absorption, i.e., the presence of the trace gases at a height of 5–10 km as well. Fortunately, there is a solution to this. The technique is called DIAL or DIfferential Absorption Lidar. Essentially, it constitutes a slight modification of the laser system. Instead of just sending out radiation at one wavelength, the laser system now generates radiation at two wavelengths: one right on resonance with the trace gas as before and one slightly off the resonance such that this second beam is not absorbed by the trace gas. The idea is that any cloud in the atmosphere will influence the two wavelengths in the exact same manner. But the trace gas only absorbs radiation at one of the wavelengths. Thus, the difference in the two signal strengths is only caused by the presence of the trace gas. The effect of clouds would show up in both components and could thus be identified as being caused by clouds or any other scatterer.

9

It is easily conceivable that these types of measurements are quite complex and many challenges had to be met. As an example we want to discuss two issues. First, it would be ideal to make these types of measurements irrespective of the time of day. However, during daytime there is a lot of background radiation due to the sunlight; it easily swamps the small fraction of backscattered light. Thus, it was necessary to develop techniques to suppress the daylight. The solution was essentially narrow bandwidth filters that suppress the very broad spectral content of the sunlight without attenuating the returning laser light. While this constitutes a technical problem, the second issue is more fundamental in nature. It is the so-called inverse problem. Despite the fact that two wavelengths are used for the DIAL system, many effects can contribute to a certain return signal. It might be clouds, water droplets, snowflakes, ice, aerosols, very small volcanic debris from an eruption, etc. Moreover, the concentration distribution of a trace gas might be more complicated than indicated in our example. Thus, the unambiguous extraction of the actual atmospheric layering from the data is actually very complex and intricate. Bohren and Huffman have compared this problem in their book [34] on absorption and scattering by small particles to the problem of a knight hunting a dragon: it is actually quite straightforward to recognize the footprint of a dragon when seeing the dragon; it is very hard to conclude what the dragon might look like from just seeing its footprints.

The DIAL technique was developed in the early 1970s and 1980s as a method of sensitive trace gas detection in the atmosphere. At first DIAL was used to monitor many air pollutants and later it was used to monitor the depletion of the ozone layer. In principle, ozone can be detected using passive remote sensing techniques aboard satellites. The disadvantage is, however, that only the cumulative concentration can be measured. Any height information is lost. While this

height information can be extracted via balloon measurements, the laser based DIAL technique provides data much faster, more reliably and up to higher altitudes than the balloon based measurements; it is the superior approach.

9.4.7 Femtosecond-Lidar Application for Influencing Weather Phenomena

As noted earlier, a short pulse has a very large spectral bandwidth; hence, more than one trace gas species could potentially be detected simultaneously. Different spectral parts of the pulse would be absorbed by different species, and one laser with sufficiently short pulses should be able to simultaneously detect an entire range of trace gases. When researchers from France and Germany tried this for the first time at a relatively short distance, the technique indeed worked. However, when they directed the laser into the atmosphere, they discovered a surprise. Instead of a well-defined absorption at particular frequencies, they observed a bright white light extending a few hundred meters into the air. By beam steering and choice of laser parameters the researchers were able to move this channel to a variety of different locations. A thorough analysis revealed that the observation was actually a plasma channel forming due to the high intensity of the pulsed laser. The laser would ionize the air, producing a plasma state of free electrons and ions. The white light then originates from the ions and electrons recombining. In principle such a plasma state should make the laser diverge rapidly due to the contributions of the free electrons to the index of refraction. However, there is a focusing effect that counterbalances it so that propagation of the plasma channel is, in fact, relatively stable over a large distance. This self-focusing of a high intensity laser is known as the Kerr-effect. For high intensities the index of refraction has a non-linear contribution proportional to the intensity of the pulse. Since laser beams usually have a profile with higher intensity in the middle, the index of refraction is higher in the center; this leads to a focusing of the beam. Consequently, the two effects cancel each other and literally form an artificial lightning strike moving up into the atmosphere. Immediately, this sparks the creativity of a physicist. What if such a laser could be used to "direct" a natural lightning strike in a thunderstorm? Instead of more or less random lightning strikes, where it hits could be selected by essentially sending the artificial lightning strike towards an area where the next natural lightning strike is suspected. First experiments already hint at this becoming a reality in the near future. In thinking even further ahead, it is conceivable that using such a plasma channel could produce rainfall from clouds. Today, clouds are seeded by silver iodine in order to improve rain precipitation. In the future, it might become possible to just use such a femtosecond laser system to make rain from a cloud.

9.4.8 Stand-Off Super-Radiant Spectroscopy

Stand-off super-radiant spectroscopy is a promising remote sensing technique that is far more sensitive than Lidar or DIAL [35]. Briefly, the idea for detection of a weak concentration of some species is to use laser pulses to provide optical pumping of the species at a distant location in such a way that it will actually send back a laser beam signal. This is achieved by sending out multiple pairs of pulses along the same path; one pulse in a pair has wavelength λ_1 and the other λ_2. There is a small time delay $\Delta\tau$ between the two pulses in a pair and if $\lambda_1 > \lambda_2$, then the λ_2 pulse must be in the lead. The idea is that due to atmospheric dispersion, the longer wavelength pulse λ_1 will move faster and catch up to the shorter wavelength pulse at a well-defined distance determined by $\Delta\tau$. The wavelengths λ_1 and λ_2 are

chosen so that when they both interact with a gas molecule of interest, the molecule will be driven into an appropriate excited state (the upper lasing level) via two photon pumping, or via some Raman process. So, for a given $\Delta\tau$, there will be a well-defined distant point where the molecules of interest will be excited and produce a gain region. By sending out a long sequence of pulse pairs in which $\Delta\tau$ steadily decreases for each pair, there will be a long sequence of these gain regions being created in a direction directly back to the observer. Basically, it is a gain swept amplifier that is lasing straight back to the observer.

This approach will make it possible to detect parts per million concentrations of dangerous gases at distances of several kilometers. An important application of the gain swept amplifier is to actually obtain backward directed lasing from the major constituents of air, N_2 and O_2 [36]. A photon from this backward lasing can combine with a photon from a forward propagating interrogation laser beam to produce a molecular excitation in some trace gas. Then, by measuring the absorption of the backward propagating laser beam as a function of the wavelength of the interrogation laser, the trace gas concentration can be quantified with great sensitivity. Finally, this backward directed lasing could be an important tool in astronomical adaptive optics [37]; it could provide an artificial guide star at any position in the sky.

9.5 Conclusions

Optics and in particular light in remote sensing has evolved in the past decades to be one of the most important applications for learning more about the environment in which we live. The variation of optical methods for remote sensing is very large; it spans the gamut from airplane and satellite based passive remote sensing platforms to active remote sensing using time-of-flight techniques. In active remote sensing, the development of the laser and subsequently the introduction of Lidar have led to particularly striking progress. Lidar can be used for a large variety of tasks that shape knowledge about our environment—ocean temperature, trace gas detection in the atmosphere, air pollution in cities, telemetry, and many other applications. Once more optics and light is demonstrating how important and fundamental it is to our lives.

References

1. Robert Bosch Tool Corporation (2015) Power tools for professionals GLR825. ► http://www.boschtools.com/Products/Tools/Pages/BoschProductDetail.aspx?pid=glr825-specs
2. Dickey JO, Bender PL et al (1994) Lunar laser ranging: a continuing legacy of the Apollo program. Science 265(5171):482–490
3. Guizzo E (2011) How Google's self-driving car works. IEEE Spectrum. ► http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/how-google-self-driving-car-works
4. Whitwam R (2014) How Google's self-driving cars detect and avoid obstacles. Extremetech. ► http://www.extremetech.com/extreme/189486-how-googles-self-driving-cars-detect-and-avoid-obstacles
5. Army Test and Evaluation Command, Aberdeen Proving Ground, MD (1969) Laser rangefinders. Ft. Belvoir Defense Technical Information Center
6. Litz B (2014) Laser rangefinders; Chapter 16 in Modern advancements in long range shooting. Applied Ballistics LLC, Cedar Springs, MI
7. Northrop Grumman Systems Corporation (2013) MARK VII handheld eyesafe laser target locator. ► http://www.northropgrumman.com/Capabilities/MarkVII/Documents/markvii.pdf
8. Owen D (2011) Stuff I like: long-distance operators. Golf Dig 62(3):70–71
9. Pappas S (2013) Legend of lost city spurs exploration, debate. Live Science. ► http://www.livescience.com/37539-legend-ciudad-blanca-lost-city.html
10. Wright CW, Brock JC (2002) EAARL: a lidar for mapping shallow coral reefs and other coastal environments. In: Proceedings of the 7th international conference on remote sensing for marine and coastal environments, Miami
11. Brock JC, Wright CW et al (2004) LIDAR optical rugosity of coral reefs in Biscayne National Park, Florida. Coral Reefs 23:48–59
12. Young HD, Freedman RA (2014) University physics with modern physics. Pearson
13. Fry ES (2012) Remote sensing of sound speed in the ocean via Brillouin scattering. In: Hou W, Arnone R (eds) Proceedings of SPIE 8372. pp 8372071–8372078
14. Hickman GD, Harding JM et al (1991) Aircraft laser sensing of sound velocity in water: Brillouin scattering. Remote Sens Environ 36:165–178
15. Guagliardo JL, Dufilho HL (1980) Range-resolved Brillouin scattering using a pulsed laser. Rev Sci Instrum 51:79–81
16. Cummins HZ, Gammon RW (1966) Rayleigh and Brillouin scattering in liquids: the Landau—Placzek ratio. J Chem Phys 44:2785–2796
17. Rouch J, Lai CC et al (1976) Brillouin scattering studies of normal and supercooled water. J Chem Phys 65:4016–4021
18. O'Connor CL, Schlupf JP (1967) Brillouin scattering studies of normal and supercooled water. J Chem Phys 47:31–38
19. Fry ES, Emery Y et al (1997) Accuracy limitations on Brillouin lidar measurements of temperature and sound speed in the ocean. Appl Opt 36:6887–6894
20. Schorstein K, Scheich G et al (2007) A fiber amplifier and an ESFADOF: developments for a transceiver in a Brillouin-LIDAR. Laser Phys 17:975–982
21. Popescu A, Walther T (2009) On an ESFADOF edge-filter for a range resolved Brillouin-lidar: the high vapor density and high pump intensity regime. Appl Phys B 98:667–675
22. Rudolf A, Walther T (2012) High-transmission excited-state Faraday anomalous dispersion optical filter edge filter based on a Halbach cylinder magnetic-field configuration, Opt Lett 37:4477–4479
23. Rudolf A, Walther T (2014) Laboratory demonstration of a Brillouin lidar to remotely measure temperature profiles of the ocean. Opt Eng 53:051407, 1–9
24. Shi J, Li G et al (2007) A lidar system based on stimulated Brillouin scattering. Appl Phys B 86:177–179
25. Shi J, Ouyang M et al (2008) A Brillouin lidar system using F–P etalon and ICCD for remote sensing of the ocean. Appl Phys B 90:569–571
26. Fry ES, Katz J et al (2002) Temperature dependence of the Brillouin linewidth in water. J Mod Opt 49:411–418
27. He X, Wei H et al (2012) Experimental measurement of bulk viscosity of water based on stimulated Brillouin scattering. Opt Commun 285:4120–4124
28. Xu J, Ren X et al (2003) Measurement of the bulk viscosity of liquid by Brillouin scattering. Appl Opt 42:6704–6709
29. National Oceangraphic Data Center (NODC), User Services Branch, NOAA/NESDIS E/OC21 (1993) Oceanographic station profile time series
30. The International Association for the Properties of Water and Steam (1997) Release on the refractive index of ordinary water substance as a function of wavelength, temperature and pressure. ► http://www.iapws.org/relguide/rindex.pdf

31. Grosso VAD (1974) New equation for the speed of sound in natural waters (with comparisons to other equations). J Acoust Soc Am 56:1084–1091
32. Fry ES, Kattawar GW et al (2000) System and method for detecting underwater objects. Texas A&M University, Patent# 6388246
33. Gong W, Dai R et al (2004) Detecting submerged objects by Brillouin scattering. Appl Phys B 79:635–639
34. Bohren CF, Huffman DR (1983) Absorption and scattering of light by small particles. Wiley, New York
35. Kocharovsky V, Cameron S et al (2005) Gain-swept superradiance applied to the stand-off detection of trace impurities in the atmosphere. Proc Natl Acad Sci U S A 102:7806–7811
36. Hemmer PR, Miles RB et al (2011) Standoff spectroscopy via remote generation of a backward-propagating laser beam. Proc Natl Acad Sci U S A 108:3130–3134
37. Wizinowich PL, Mignant DL et al (2006) The W. M. Keck observatory laser guide star adaptive optics system: overview. Publ Astron Soc Pac 118:297–309

Optics in Nanotechnology

Munir H. Nayfeh

M.H. Nayfeh (✉)
Department of Physics, University of Illinois at Urbana-Champaign, 1110 W. Green Street, Urbana, IL 61801, USA
e-mail: m-nayfeh@illinois.edu

M.D. Al-Amri et al. (eds.), *Optics in Our Time*, DOI 10.1007/978-3-319-31903-2_10

10.1 Introduction

Optics is one of the most important branches of physics. It involves the study of the behavior and properties of light in vacuum as well as the study of its interactions with matter in the gas, liquid, and solid states [1]. Moreover, the field encompasses the construction of instruments that use or detect light that may serve many other fields, over a wide range of the electromagnetic waves from UV to infrared light.

Nanotechnology or nanoscience and technology, on the other hand, aims at construction, understanding, and putting to use ultrasmall particles [2–4]. Miniaturization of all types of matter including dielectrics, metals, semiconductors, polymers, etc., affords interesting novel properties, especially optical properties. The novel size regime is intermediate between the largest molecules and 100 nm. In this regime, phenomenon may not be as predictable as those observed at larger scales. Using nanoparticles as building blocks to construct advanced devices that exploit their novel properties is at the heart of nanotechnology.

It is to be noted that naturally existing colloids, micelles, polymer molecules, and phase-separated regions in block copolymers, for example, fall in this size regime. More recently, naturally unknown but interesting classes of nanostructures such as carbon nanotubes [5], silicon nanoparticles [6, 7], metal nanoparticles and nanorods [8, 9], and compound semiconductor quantum dots [10] have been designed and fabricated. The application of such extremely small particles can find applications across all fields [4, 11]. In this article we will focus on how miniaturization down to the nanoscale regime impacts the behavior, properties, and interactions of light with matter, especially metals and semiconductors, and how it enables novel advanced devices with application in electronics, photonics, energy and lighting, and biomedicine.

10.2 Optics in Nanometals: Nature of Interaction of Light with Metal

In a dielectric, all electrons are bound in atoms; and each atom interacts with light individually through the interaction of a single bound electron. The total effect in a sample is simply the sum of the individual atomic responses. In a metal solid, some electrons are bound to atoms and others are not bound in specific atoms or ions. Upon interaction with light, two processes can take place in metal. In interband transition, bound electrons, i.e., electrons in the valence band can be promoted due to light absorption to an empty level in the conduction band where they become not bound to a specific ion. Interband absorption forms a significant loss mechanism in metal at optical frequencies. On the other hand, in a metal solid, electrons that are already in the conduction band form a see or cloud of electrons not bound to specific atoms or ions. The cloud can interact with and move under an external electric force collectively at the same time.

10.2.1 Plasma Model

Figure 10.1 (left) shows a schematic of the lattice of bulk metal where electrons are subjected to a force created by the electric field of light. Free electrons move opposite to the direction of the electric field while positive ions are stationary, thus causing a shift between the center of distribution of the negative charge and the positive charge. In the absence of the force the centers of the two distributions coincide. Figure 10.1 (right) shows the same situation for a very small sphere of metal, namely a nanoparticle.

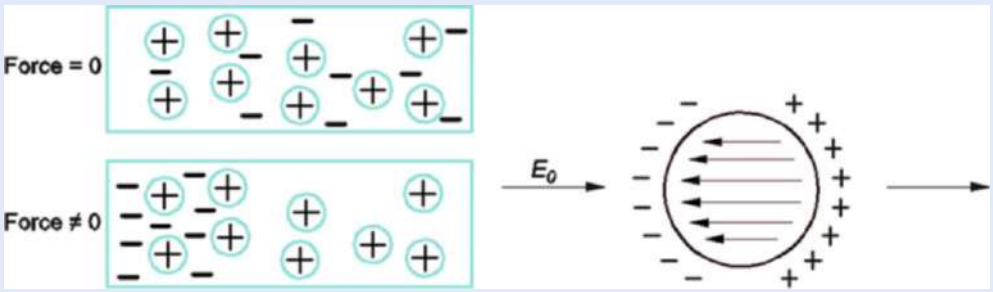


Fig. 10.1 Sketch of metal lattice subjected to a force created by an electric field. (*Left*) bulk and (*Right*) nanoparticle

Electrons in a small sphere are described by a simplified model of a gas of free electrons that moves against a fixed background of positive ion cores [12–14]. This has been called a *plasma model*. In the model, details of the lattice potential and electron–electron interactions are not included explicitly; rather they are simply incorporated into the effective optical mass m of each electron. Under irradiation, the negative electron's see or cloud gets pushed back and forth relative to the fixed background of positive ion at the frequency of oscillation in the electric field of the electromagnetic light wave. In addition, the cloud experiences an effective drag force due to collisions with the positive ion core which tends to slow it down. According to the plasma model, electrons oscillate and their motion is damped with a characteristic collision frequency γ as high as 10^{14} per second (100 THz) (corresponding to the frequency of infrared light). The time elapsed between two consecutive collisions, $\tau = 1/\gamma = 10^{-14}$ s, is known as the relaxation time of the free electron gas. The balance of these two forces gives the response or what is called the real part of the dielectric function of the cloud, $\epsilon(\omega)$, to the light wave:

$$\varepsilon(\omega) = 1 - \frac{ne^2}{\varepsilon_0 m\omega^2}$$

Where n is the density of electrons in the cloud, e and m are the charge and the effective mass of a single of electron, ϵ_0 is the dielectric function of vacuum, and ω is the frequency of the incoming light wave. It is customary to define a useful quantity called plasma frequency or plasmon frequency ω_p in terms of the following group of constants:

$$\omega_p = \sqrt{\frac{ne^2}{\varepsilon_0 m}}$$

In terms of the plasma frequency, the dielectric function takes the simple form $1 - \omega^2/\omega_p{}^2$. For a metal with a free electron density of $10^{23}/\text{cm}^3$, ω_p corresponds to the frequency of ultraviolet light (plasmon energy $\hbar\omega_p$ ~10 eV) which is 100 or even 1000 times larger than the relaxation frequency γ.

Figure 10.2 plots $\epsilon(\omega)$ as a function of frequency. At resonance with the plasma frequency ($\omega = \omega_p$), $\epsilon(\omega)$ drops to zero. This is a plasma oscillation resonance. For light with a frequency below the plasma frequency, the dielectric function is negative and the light cannot penetrate the sample, rather it is totally reflected. Above the plasma frequency the light waves penetrate the sample as sketched in the figure.

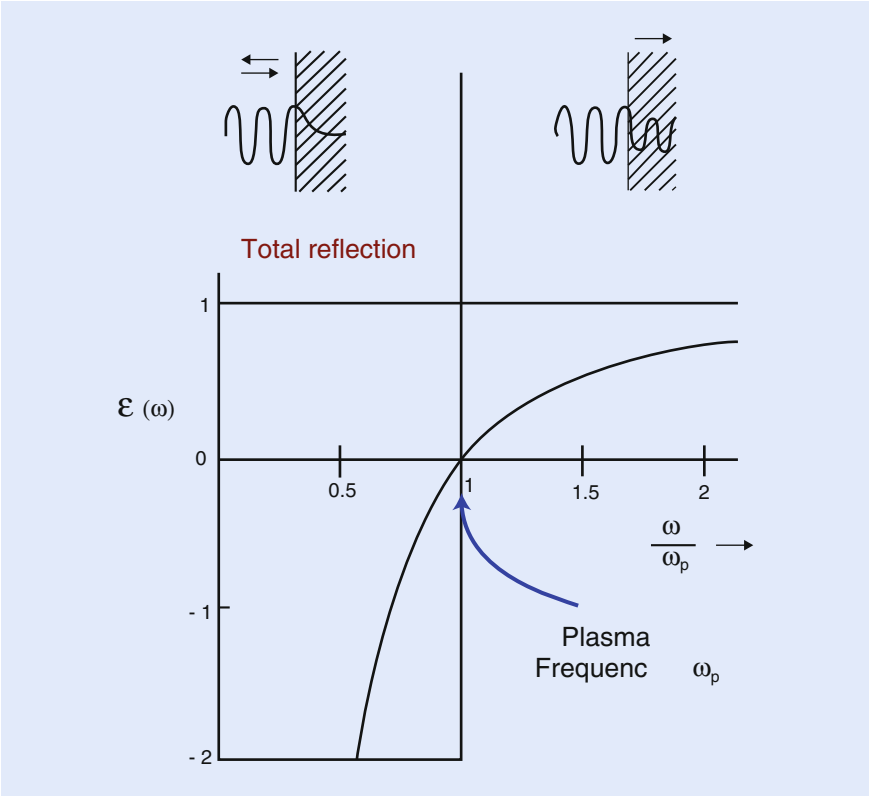


Fig. 10.2 Plot of the dielectric function $\epsilon(\omega)$ as a function of frequency for a metal. $\epsilon(\omega)$ drops to zero at resonance with the plasma frequency ($\omega = \omega_p$) (Image from ▶ https://www.coursehero.com/file/10546609/Plasmonics/)

10.2.2 Miniaturized Metal: Subwavelength Concentration of Light

We now discuss how the interaction of metal with light manifest itself as we reduce the size of the metal sample from large bulk samples to small nanoparticles. We will compare three kinds of structures sketched in Fig. 10.3: bulk, which is labeled as three-dimensional (3D); a sheet or quantum well, which is labeled as two-dimensional (2D); quantum wire, which is labeled as one-dimensional (1D); and nanoparticle or quantum dot, which is labeled as zero-dimensional (0D).

Bulk Material (3D)

Because free conduction electrons can move over very large distances in bulk metal objects, electrons on average do not oscillate against a certain localized ions with a restoring force. Instead the motion is actually similar to a mass being dragged in a viscous fluid. If the light has a frequency above the plasma frequency (in the ultraviolet (UV) range for metals), electrons will not even oscillate and the light will simply be transmitted or absorbed in interband transitions of the metal. If the light has a frequency below UV, electrons will oscillate out of phase with the incident light, causing a strong reflection. At the plasma frequency, the dielectric function is 0.

Thin Film or Sheet (2D)

When bulk metal is shrunk to a thin film of a few nm thicknesses, electrons cannot move very far in the direction normal to the film. In this case there will be electron oscillations upon light exposure; but they will only exist at the surface of the film.

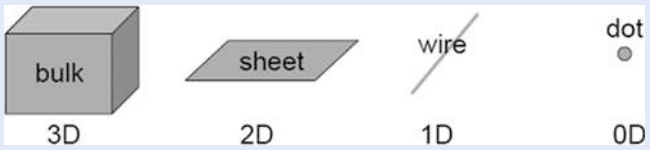


Fig. 10.3 Schematic of four kinds of structures: bulk, a sheet or quantum well, quantum wire, and nanoparticle or quantum dot

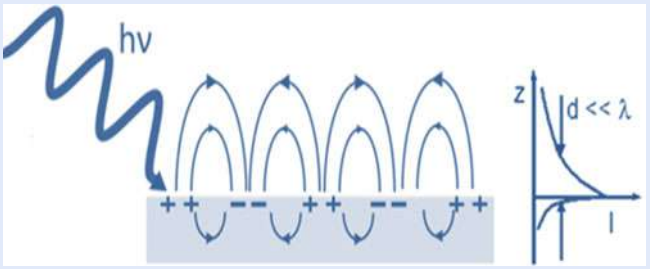


Fig. 10.4 Schematic of light interaction with a few nm thin metal films (*left*) shows electron oscillations at the surface of the film, accompanied with propagation of charge waves along the film. (Adapted from ► https://en.wikipedia.org/wiki/Surface_plasmon). (*Right*) Sketch of the field intensity with distance from the surface

This is accompanied with propagation of charge waves along the film. This is known as surface plasmon polaritons (SPPs) as sketched in Fig. 10.4 (left). In fact, detailed Maxwell's theory shows that the surface waves can propagate along the surface with a broad spectrum of frequencies from 0 up to $\omega_p/\sqrt{2}$, where ω_p is the plasma frequency defined above. At $\omega_p/\sqrt{2}$ the dielectric function is -1 with a charge-wave wavelength shorter than the wavelength of the incident light. The combined effect is a mixed or hybrid light-electron-wave-state. This results in intense light–matter interactions with unprecedented optical response. As shown in Fig. 10.4 (right), the intense optical field resulting from the hybrid field is a local one, extending outside the thin film into the dielectric only few nanometers, a distance much less than the wavelength of the incident light ($d \ll \lambda$). Thus visible light, which has a wavelength of approximately half a micrometer, can be concentrated by a factor of nearly 100 to travel through metal films just a few nanometers (nm) thick.

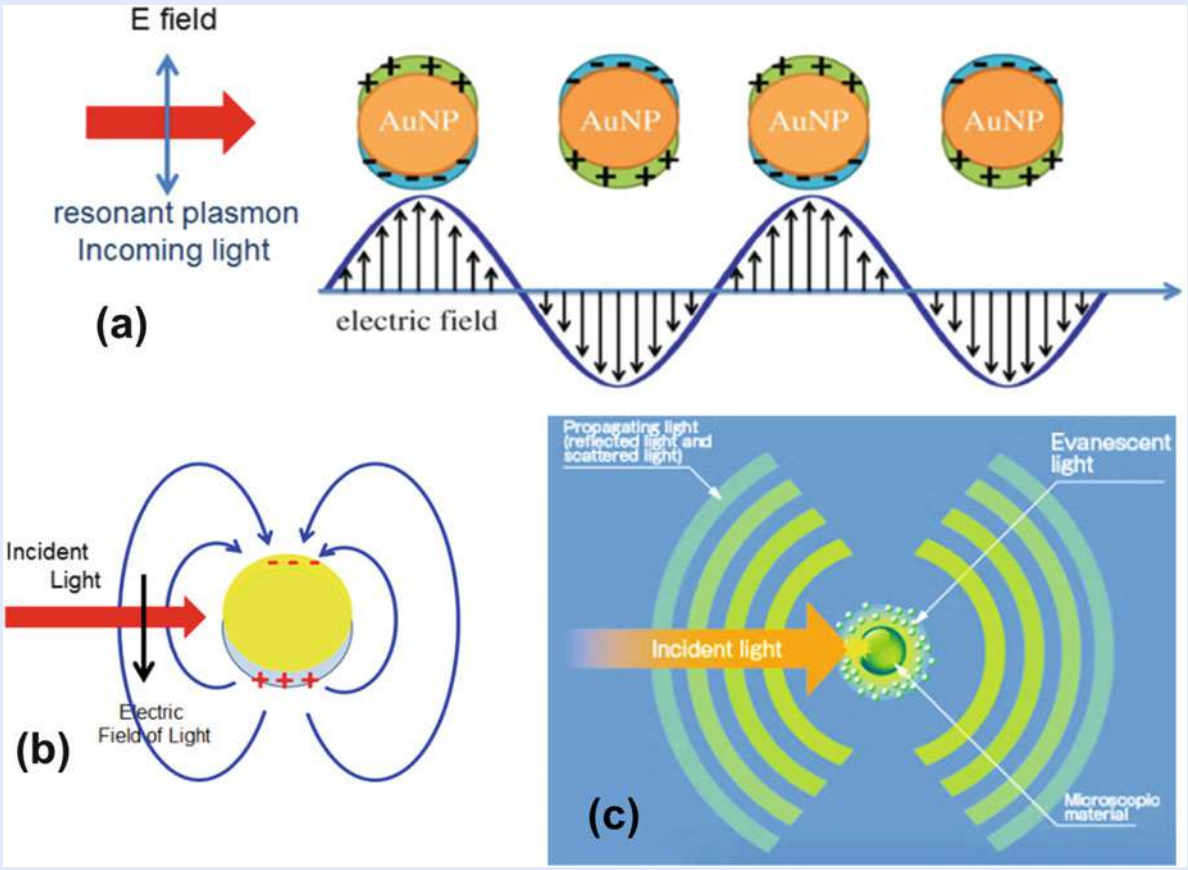
Nanowire (1D)

A variety of metal nanowires have been fabricated and studied. They may come as solid nanowire of metal with a surrounding dielectric jacket, i.e., glass or air. For example, chemically prepared silver nanowires ~100 nm diameters were fabricated [15]. The wires were found to support surface plasmon modes propagating along the wires. The wavelength of the propagating charge wave is shortened to about half the wavelength of exciting light. The propagation length of SPP is about 10 μm. The reflectivity at both ends of the wire is about 25 %. Those characteristics are sufficient for the wires to be used as an optical instrument, namely surface plasmon Fabry–Perot resonators. Wires can also be in the form of a cylindrical shell of metal. The shell may have an inner dielectric jacket, such as glass (silicon oxide). The bore of the system may be filled with a semiconductor, such as Si or CdSe. More complicated variations of these architectures have been fabricated to accommodate ports for light entry or extraction. Semiconductor wires with

hemispheres of metal on both ends have also been fabricated. The schemes involving hybrid semiconductor-plasmonic (metal) architectures will be discussed in ◘ Sect. 10.4.1 on integrating optics with electronics.

Nanoparticles/Dot (0D)

Bulk is now shrunk in three dimensions to form particles, such that the dimension of the particle is less than the wavelength of incident light. Because the size of the particle is small compared to the wavelength the incident electric field will be constant across the nanoparticle, inducing a uniform displacement of the electron density, making the electron motion in phase (collective cloud motion). Because electrons are confined within the small particle, they will oscillate while being accompanied by a strong restoring force from specific positive ionic core background (◘ Fig. 10.5a). The restoring force leads to non-propagating (also called evanescent) collective oscillation of the surface cloud with a characteristic oscillation frequency similar to a simple harmonic oscillator. This is unlike bulk which does not have a specific oscillation frequency. ◘ Figure 10.5b shows representative field lines, effectively resembling those of an oscillating electric dipole. Another



◘ **Fig. 10.5** Schematic of the interaction of a gold nanoparticle with linearly polarized light beam. (**a**) Electrons oscillate while being accompanied by a strong restoring force from specific positive ionic core background (Adapted from E. Yasun, H. Kang, H. Erdal, S. Cansiz, I. Ocsoy, Y-F. Huang, W. Tan, Cancer cell sensing and therapy using affinity tag-conjugated gold nanorods, Interface Focus 3 (2013) 0006: ▶ http://dx.doi.org/10.1098/rsfs.2013.0006). (**b**) Representative field lines, effectively resembling those of an oscillating electric dipole. (**c**) Region where the non-propagating (evanescent) light is localized, related only to the size of the nanoparticle, not to the wavelength (Image adapted from [16])

striking effect is that the collective oscillations lead to a large absorption and scattering cross section, as well as an amplified local optical electromagnetic field. The electric field has its maximum strength just outside the nanoparticle and drops rapidly with distance extending only to ~30 nm. For small particles less than ~15 nm, light absorption dominates; whereas for nanoparticles greater than ~15 nm scattering dominates. In fact, detailed Maxwell's theory shows that surface waves can be resonantly sustained with frequency at $\omega_p/\sqrt{3}$, which corresponds to dielectric function of -2. We should stress because the evanescent light is non-propagating it is localized within the approximate radius of the particle, even if the size of that particle is much smaller than the incident light wavelength (Fig. 10.5c). The region where the non-propagating evanescent light is localized is related only to the object's size and not to the wavelength, which can be considered as light with no diffraction limit. The nanoparticle practically confines light into super intense "hot-spot."

10.2.3 Miniaturization-Induced Coloration of Metals

Figure 10.6 (top left) shows a large block of gold under illumination by white light. When light strikes the large block, the red and green components are reflected; while the blue component is absorbed and scattered. Interband transitions of bulk gold give it a yellow color. But a metallic luster is also added

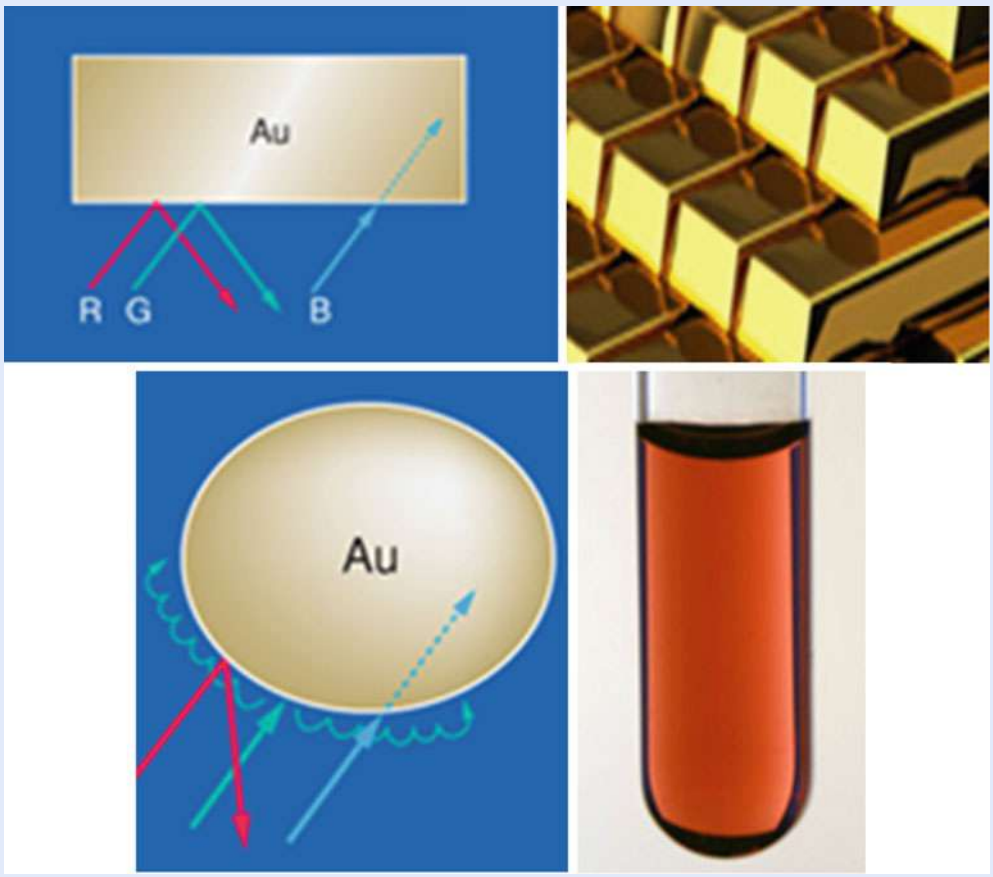


Fig. 10.6 Schematic of the interaction of gold with white light, showing the individual responses to red, green, and blue (RGB) components, as well as the resulting color as seen by the naked eye (*top row*) large block of gold and (*bottom row*) colloid of gold nanoparticles. (Adapted from [16])

to the yellow color, which appears as a golden color (Fig. 10.6 top right). Charge oscillations in bulk do not have a specific frequency to cause specific coloration.

When white light strikes, on the other hand, the nanoparticle shown in Fig. 10.6 (bottom left), the color will be different. This is due to a specific plasmon resonance due to a confined collective oscillation of electrons. A plasmon resonance in gold nanoparticles occurs over a thin slice of light at a frequency in the green. Thus not all of the components of incoming white light resonate and are absorbed. Among the RGB components, only the G (green) component resonates with the electrons and is absorbed in the gold nanoparticle. The B (blue) light is absorbed and scattered, and only the remaining R (red) component gets reflected or passes through. This is the reason that stained glass mixed with gold nanoparticles appears red to the naked eye. Figure 10.6 (bottom right) demonstrates that, when light strikes a gold nanoparticle colloid, only the red color is reflected or passes through [16].

10.2.4 Plasmonic Lenses

Having a negative refractive index is the basic principle behind plasmonic lenses. This is a very common property for noble metal at specific frequency. There are two types of such lenses. One type is based on confinement enhancement. Another is based on transmission enhancement of evanescent waves. In this section we present examples of metal-based lenses.

Confinement-Based Lensing

We discuss two configurations for confinement-based lensing: a continuous conic waveguide concentrator and discontinuous chain of nanoindentation (particles). (1) Consider the hollow cone of metal shown in Fig. 10.7 (left). The radius of the cone gradually decreases from 50 nm to 2 nm, for example. When light strikes the cone and a plasmonic resonance is excited at the opening of the cone, a surface plasmonic polariton (SPP) propagates towards the tip. This causes accumulation of energy resulting in a giant local field at the tip, as shown in Fig. 10.7 (left). Figure 10.7 (right) displays the amplitudes of the local optical field in the cross section of the cone for the normal and longitudinal (with respect to the axis) components of the optical electric field. The magnitude of the field grows significantly as the oscillations approach the tip. The transverse x component grows by an order of magnitude, while the longitudinal z component, which is very small far from the tip, grows relatively much stronger at the tip. This causes the local field to increase by nearly 3 orders of magnitude in intensity and four orders in energy

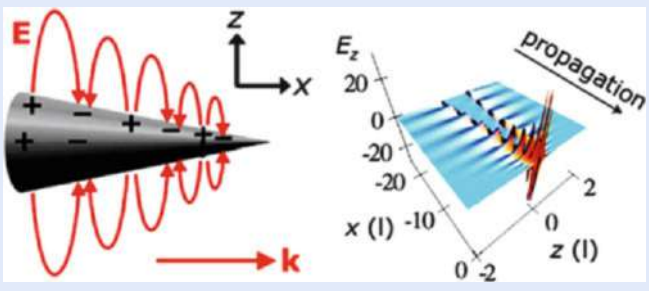


Fig. 10.7 (*Left*) Geometry of a conic nanoplasmonic waveguide showing propagation of a charge oscillation wave. (*Right*) Snapshot of instantaneous E_z fields in the longitudinal cross section (*xz*) plane, normalized to far-zone (excitation) field. (Adapted from "Stockman MI (2004) Phy Rev Lett 93: 137404")

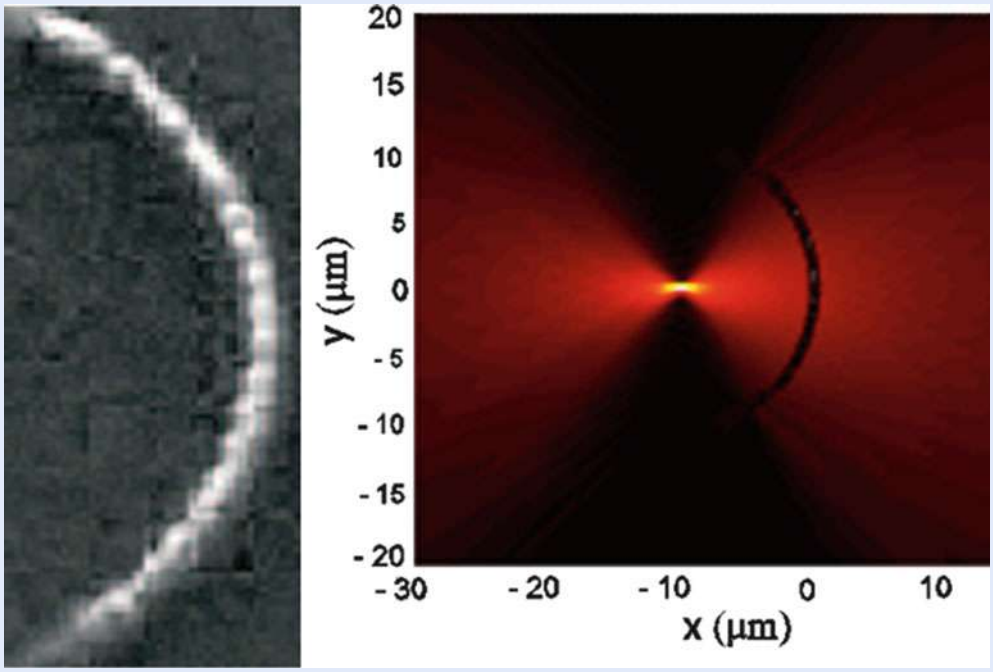


Fig. 10.8 Focusing of light by a curved chain of spheroid gold nanoparticles indented in a thin gold film (*left*) top view image of the chain using a scanning electron microscope (SEM), (*right*) Magnitude of scattered electric field calculated above the gold surface. The illuminating beam has a wavelength of 800 nm and is incident perpendicular to the gold surface and polarized along *x*-direction (Adapted from "Evlyukhin AB, et al. (2007) Opt Exp 15: 16667–16680")

density. It should be noted that the propagation is slowed down and asymptotically stopped when it approaches the tip. It never actually reaches the tip (the travel time to the tip is logarithmically divergent) [17]. The cone as such may represent a tapered plasmonic waveguide. (2) Consider a chain of nanoparticles shown in Fig. 10.8 (left). Fabrication starts with a thin gold film. Nanoindentations (nanoparticles) are made on the film in a parabolic chain configuration. The radius of curvature of the chain is 10 μm. The particle diameter and inter-particle distance are ~350 and 850 nm, respectively, while the particle height is 300 nm. When light, in the wavelength range of 700–860 nm, for example, is normally incident on the thin film on the right of the chain, charge oscillations are induced on the surface of the film which travel or propagate to the left towards the chain (surface plasmon polariton or SPP). When the waves hit the gold nanoparticles they get focused to a submicron spot, as shown in Fig. 10.8 (right) [18].

Transmission-Based Lensing

Another concept of lensing involves transmission through holes or slits. An array of holes is drilled in an opaque metallic film in an arc formation, as shown in Fig. 10.9a. The diameter of the holes is of nanoscale. For a plane wave incident upon such a structure, the phase shift experienced by light as it passes through each individual hole is sensitive to either the length, width, or even the materials inside the hole. With the adjustment of the properties of individual holes, it becomes possible to achieve a focusing action [19].

A slit-based lens is built by first depositing a 400 nm thick flat gold film on silicon oxide substrate. Then an ion beam is used to mill 13 rectangular slits in the thin film, as shown in the left panel of Fig. 10.9b. The slits increase in widths from 80 nm at the center slit to 150 nm on the left or right end. Figure 10.9c

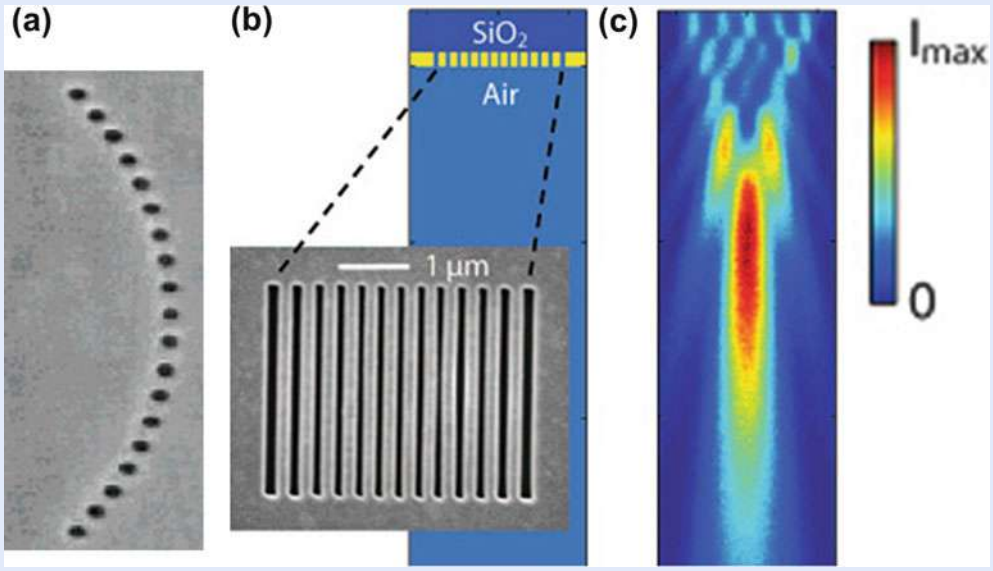


Fig. 10.9 Focusing of light by transmission through holes or slits. (**a**) Chain of holes in a metal plate. (**b**) Nanoscale slit array drilled in gold film on a fused silica substrate (*dark blue*). The film is 400 nm thick (*yellow*). The air slits are different in widths (80–150 nm) (*light blue*). The *inset* shows a scanning electron micrograph of the structure as viewed from the air-side. (**c**) Focusing pattern measured by confocal scanning optical microscopy (CSOM) (Adapted from "Lieven Verslegers et al., Nano Lett. 9, 235–238 (2009)")

gives the measured field intensity in a cross section through the center of the slits (along the *x*-direction). The measurement demonstrates focusing of the wave. Thus this configuration acts as a far-field cylindrical lens for light at optical frequencies [20].

10.2.5 Metamaterials: Negative Refractive Index

The above discussion shows that metal dielectric interface, with features smaller than the wavelength of light separated by distances smaller than the wavelength of light, is very special. At plasmon resonances they have a negative index. In fact those may be considered as a class of more general material called metamaterials that exhibit properties beyond those found in nature. In 1967 [21] Victor Veselago theorized that material with negative refractive index would exhibit optical properties opposite to those of dielectrics, such as glass or air. Contrary to dielectrics, when light propagates in metamaterial (1) light refracts on the other side of the normal to the interface, that is, energy is transported in a direction opposite to the dielectric case, as shown in Fig. 10.10; (2) light produces negative pressure, which pulls metamaterial towards it instead of positive pressure which pushes away as in conventional material. The basic principle behind all of the opposite effects is closely related to the above plasmonic effect. They are due to the collective interaction of the light with the electron clouds at the surface of the conductor. This photon–plasmon interaction generates intense, localized optical fields. The waves are confined to the interface between metal and insulator. This narrow channel serves as a transformative guide that, in effect, traps and compresses the wavelength of incoming light to a fraction of its original value.

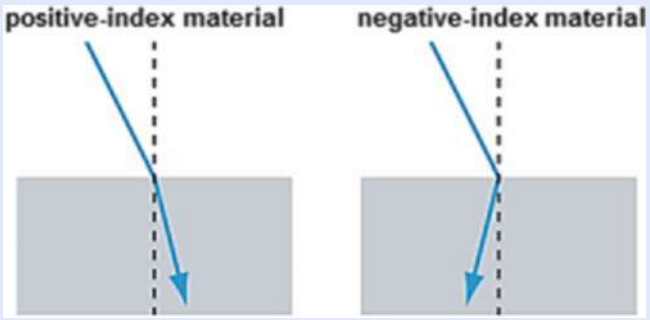


Fig. 10.10 Schematic of light refraction at a plane interface of dielectric—dielectric interface and at a dielectric—metamaterial (negative index) interface

10.2.6 Heat Loss: Are Plasmonic-Based Devices Practical?

One pivotal problem inherit in plasmonic technology is heat loss. Because plasmonic devices involve light interaction with metal, it necessarily involves energy loss due to heat dissipation. Metals are plagued by large losses due to strong electronic interband absorption, especially in the visible and UV spectral ranges. Among all metals, noble metals of gold and silver show the least energy loss. That is why gold and silver are used in most common plasmonics. Even in gold and silver, losses in the optical range including near- and mid-infrared (IR) regions are effectively still too high to make practical plasmon-based devices. Therefore, the search is continuing to find different approaches, such as creation of alloys and composites to make plasmonic materials that exhibit lower losses [22].

10.3 Optics in Nanosemiconductors

The interaction of semiconductor nanocrystals with light is fundamentally different from their bulk counterpart. Quantum dots, which were first discovered in 1980, are tiny particles or nanocrystals of a semiconducting material with diameters in the range of 2–10 nm. In this size regime, additional quantum effects start to play an important role that can significantly alter some properties of the original material, such as optical activity. The most apparent result of this is that nanocrystals can fluoresce in distinctive colors determined by the size of the particles.

10.3.1 Bandgap and Excitons

Unlike metals, semiconductors do not electrically conduct, i.e., have no free electrons in the conduction band. All of the electrons are bound to atoms, i.e., they are in the valence electronic bands. This is because in metals the conduction and valence bands overlap but in semiconductors the edges of bands are separated by an energy gap called bandgap, as shown in Fig. 10.11a [23].

In semiconductors an external agent such as external light is needed to impart enough energy to a bound electron to overcome the bandgap energy, as shown in Fig. 10.11b. When this takes place the electron is placed in the conduction band accompanied with its separation from the positive charge. Separation of the electron from the positive charge leaves behind in the valence band what is called

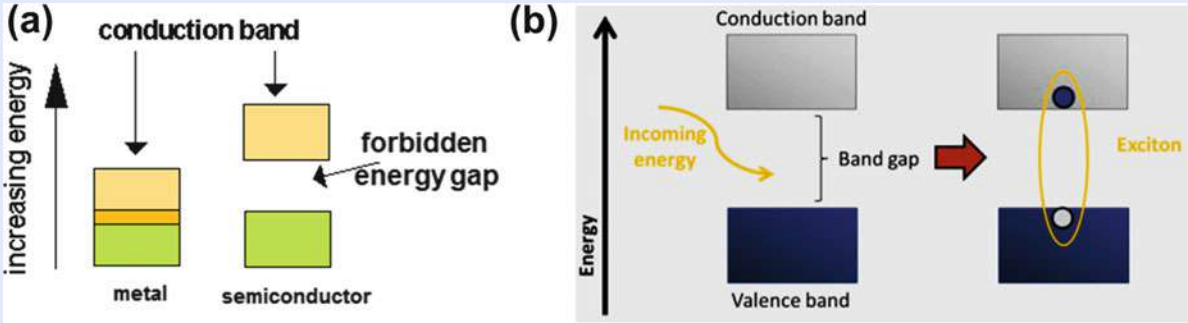


Fig. 10.11 (**a**) Conduction and valence bands for metal and semiconductor (Adapted from ▶ http://webs.mn.catholic.edu.au/physics/emery/hsc_ideas_implementation.htm). (**b**) Light excitation of semiconductor elevating an electron from the valence band to the conduction band creating an exciton (Adapted from [23])

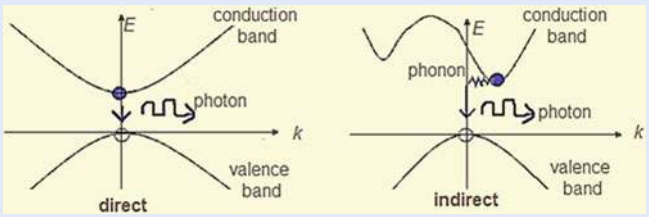


Fig. 10.12 Schematic of the band structure showing the conduction and valence bands for direct bandgap material and indirect bandgap material

a positive hole. The electron as well as the positive hole are free to move away from their original common site; but they move together as an electron–hole pair or exciton, with the electron orbiting around the hole at an average distance or exciton Bohr radius. In silicon, for example, the Bohr radius is 4.2 nm. The exciton structure is actually an "atom" which in some respect resembles the structure of a hydrogen atom, but 100–200-fold larger. Because of the large size, excitons are less bound and more fragile.

10.3.2 Direct and Indirect Bandgap Materials

The interaction (absorption or emission) of light with matter must satisfy the conservation of energy and momentum. The ability of meeting these conservations laws depends on the type of material. Two classes of semiconductors are interesting in this regard: indirect bandgap materials such as silicon and direct bandgap semiconductor such as CdSe (see Fig. 10.12). The bandgap is called "direct" if the momentum of electrons and holes is the same in both the conduction band and the valence band. In this case an electron can directly emit a photon, which conserves energy while the momentum is automatically conserved, as shown in Fig. 10.12 (left). In an "indirect" gap material, the momentum of electron and hole are not the same; hence a photon cannot be readily emitted because, to conserve momentum, the electron must pass through an intermediate state and transfer momentum in the form of phonon quanta to the crystal lattice, as shown in Fig. 10.12 (right). The probability of coincidence of three particles: electron,

hole, and phonon with specific properties is significantly lower than the coincidence of only two such particles (electron and hole). Thus, the probability of emission of a photon is much lower in indirect bandgap semiconductors than in direct bandgap ones. Bulk silicon is therefore a very poor light emitter, while in most cases, the direct bandgap semiconductors are good light emitters. Direct bandgap material is therefore used in light emitting devices, while indirect bandgap material is used in electronic devices.

10.3.3 Enhancing and Blue Shifting of Luminescence by Quantum Confinement

As we shrink the size of direct semiconductor material, novel optical properties begin to emerge as we approach a certain characteristic size scale. The characteristic scale is the electron–hole distance (the Bohr radius) of the material. For example, for CdSe the Bohr radius is 5.6 nm. In this regime, the interaction of light with the material gets modified because quantum quantization of the energy levels of the electron as well as that of the hole according to Pauli's exclusion principle becomes important. A simplified treatment considers the energy of the electron and the hole as the energy of the charge in an infinite well (see Fig. 10.13). The top well is for an electron and the bottom inverted well is for a hole.

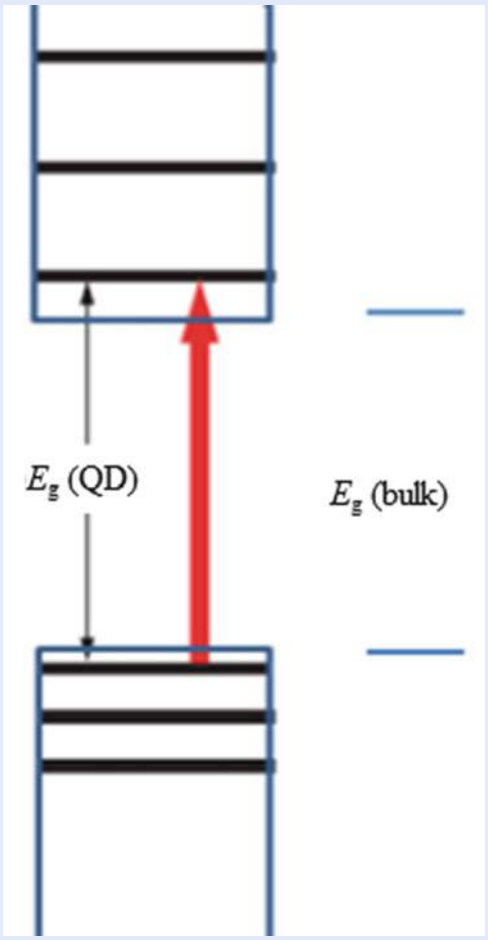


Fig. 10.13 Simplified infinite well model of the energy of the electron and the hole in a quantum dot. The top well is for an electron and the bottom inverted well is for a hole

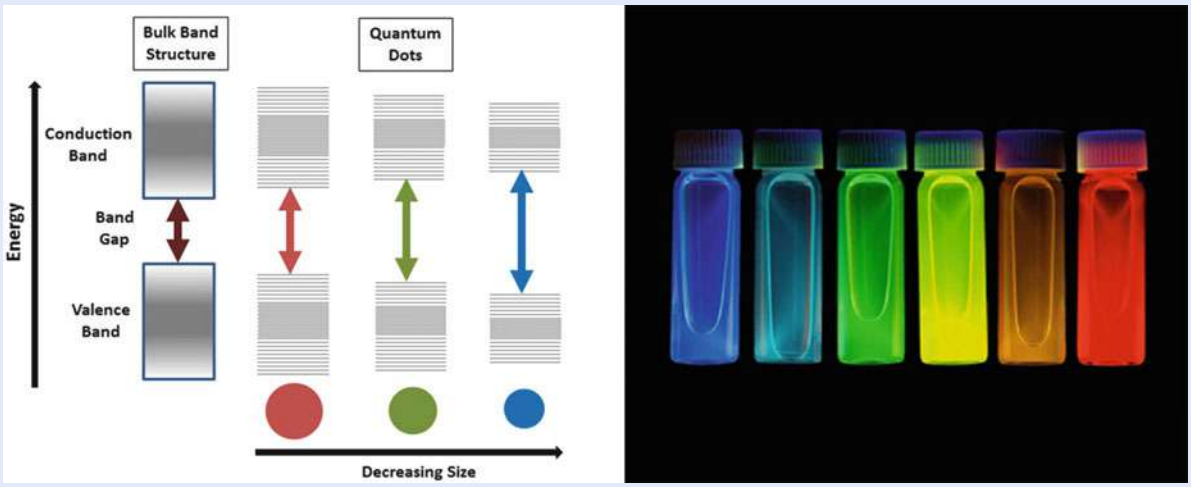


Fig. 10.14 (*Left*) The conduction and valence bands with decreasing size of a semiconductor nanoparticle (quantum dot), showing splitting of energy levels due to the quantum confinement effect. The bandgap increases with decrease in size of the nanocrystal (Images from (▶ http://www.sigmaaldrich.com/materials-science/nanomaterials/quantum-dots.html)). (*Right*) Vials of quantum dot colloids of increasing average size from left to right emitting light with color from blue to red respectively (Image from ▶ http://nanocluster.mit.edu/research.php [24])

10

This model results in three effects. First the energy levels become discrete according to: $E_n = (h^2\ n^2)/(8\ m_c\ R^2)$ where n is an integer designating energy levels, h is Planck's constant, m_c is the effective mass of the electron and hole system, and R is the radius of the quantum dot (particle). In this simplified model, the bandgap of the material widens to $E_g = E_{g0} + (h^2)/(8\ m_c\ R^2)$. Second, this result shows that as the particle decreases in size, the bandgap energy increases, as shown in Fig. 10.14 (left). More energy is then needed to excite the dot, and concurrently, more energy is released when the crystal returns to its ground state, resulting in a color shift towards blue in the emitted light. Third, in addition to the shift towards the blue the emission becomes stronger due to the discrete nature of the levels, making the nanoparticles much brighter than bulk [25]. As a result of this phenomenon, quantum dots can emit any color of light from the same material simply by changing the dot size. Additionally, with control over the size of the nanocrystals, quantum dots can be tuned during manufacturing to emit any color of light, as shown in Fig. 10.14 (right).

10.3.4 Making Silicon Glow: Quantum Confinement

Silicon is the eighth most common element in the universe by mass, but very rarely occurs as the pure free element in nature. Monocrystalline silicon, manufactured from sand using sophisticated technology, is the backbone of the microelectronics industry. However silicon is the dullest material with regard to its optical activity because it belongs to the class of indirect bandgap semiconductors. Figure 10.15 shows the energy-momentum diagram for silicon. Light emission in silicon is highly improbable because it requires both conservation of energy, which is satisfied by the emission of the appropriate photon energy at 1.1 eV, and conservation of momentum, which is satisfied by emission of a specific phonon (vibration of the crystal) of 55 meV energy. Processes that require simultaneous emission of light and vibrations are highly unlikely and take a long time to happen, providing emission lifetimes on the order of milliseconds. In direct

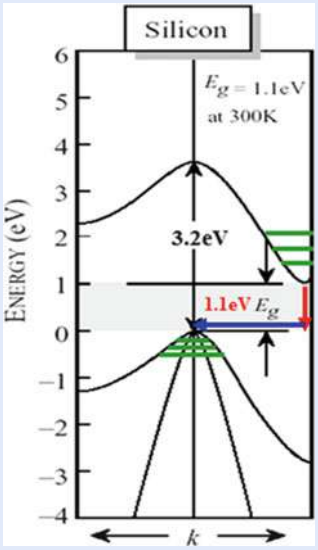


Fig. 10.15 The energy–momentum diagram of the band structure of silicon. Emission of light requires both conservation of energy, which is satisfied by the emission of the appropriate photon energy at 1.1 eV (*red vertical arrow*), and conservation of momentum, which is satisfied by emission of a specific phonon (vibration of the crystal) of 55 meV energy (*horizontal blue arrow*). The *upper set of green lines* are discrete levels due to confinement of an electron and the *lower set of green lines* are due to the confinement of a hole

semiconductors, light emission proceeds readily with a lifetime of nanoseconds as the momentum is automatically satisfied without vibration and emission of phonons.

Recent developments enabled by nanotechnology are starting to change the picture. Some silicon nanostructures provide interactions with light now approach or even exceed the performance of equivalent direct bandgap materials, which promise to take silicon into the realm of optics. A significant deviation from bulk properties was found in 1990 when L. T. Canham noticed visible photoluminescence in porous silicon he produced via chemical etching of bulk silicon in HF acid [26].

But porous silicon is just an interconnected nanoscale network of silicon skeletal (sponge-like structure). The first report of micro- and nanoparticles prepared from porous Si came from the Sailor group at University of California in San Diego [27]. Strong ultrasound was used to shatter porous silicon into micro- and nanoparticles. But the particles do not have specific configuration or uniformity because silicon is a hard material and the particles are basically a result of mechanical stress and shattering of the interconnected skeletal backbone. In 1997, a new self-limiting etching procedure was developed by the Nayfeh's group at the University of Illinois which produced on the silicon wafer disconnected individual spherical nanoparticles of preferred or magic configuration or sizes, which can be softly retrieved and stored in a liquid of choice [28–32]. The particles are protected by mono-hydride coating and can be produced in commercial amounts and stored for later use for many years. The smallest of these particles is 1 nm and fluoresces in the blue. Other sizes include 1.7 nm, fluorescing in the green; 2.15 nm fluorescing in the yellow–orange; and 2.9 nm fluorescing in the orange–red. Figure 10.16 (left) shows the luminous of colloids of 1 and 2.9 nm particles in alcohol using excitation at a wavelength of 365 nm. The 1-nm Si nanoclusters are amenable to testing and accurate first principle simulations because they consist of a manageable number of Si and H atoms and are produced in macroscopic

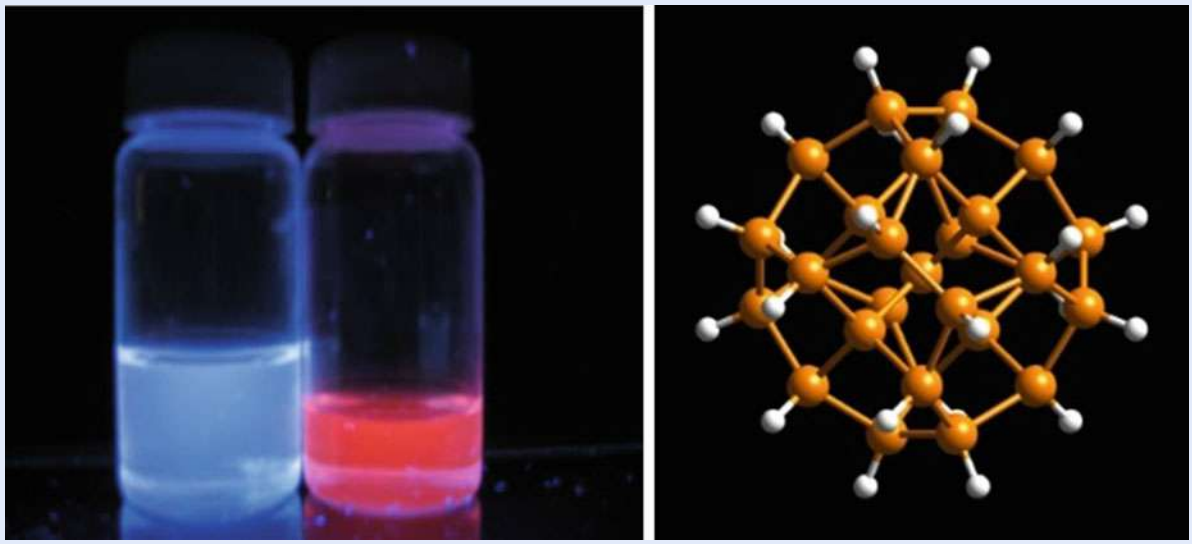


Fig. 10.16 (*Left*) Blue and red luminescence of colloids of 1 and 2.9 nm particles in alcohol, respectively, with excitation at a wavelength of 365 nm. (*Right*) A prototype structure of 1-nm particle has a configuration of $Si_{29}H_{24}$. Silicon atoms in *orange*. Hydrogen atoms in *white*

10

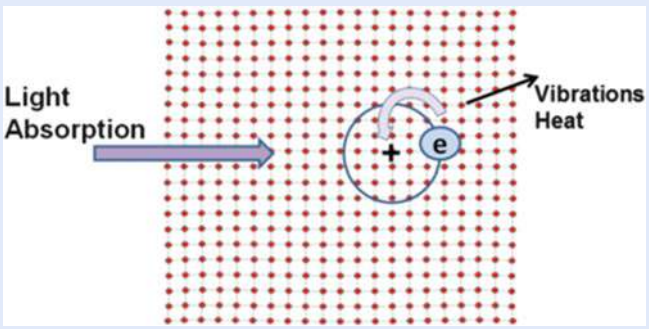


Fig. 10.17 Response of bulk silicon to light. Exciton forms followed by recombination to produce vibrations and heat before it can emit a photon

amounts. Those studies showed that the 1-nm particle simulated by Lubos Mitas has a configuration of a super molecule $Si_{29}H_{24}$, as shown in Fig. 10.16 (right) [33–37].

One important characteristic of these particles is that the energy and momentum conservation rules governing light interaction get modified. In bulk silicon crystals or large particles, photo produced excitons move freely in all directions. Because it is not easy to conserve the momentum in indirect gap material when light is emitted, they recombine again before light is emitted with energy turning into vibrations and heat imparted to the crystal, as sketched in Fig. 10.17.

When the particle is reduced to a size smaller than the exciton radius, especially in the sub 3-nm regime, the particles become less rigid allowing the silicon atoms to move, relax, or adjust especially those on the surface. Moreover, excitons get more strongly confined spatially, causing the momentum of the electron and hole to spread out appreciably according to the Heisenberg uncertainty principle, such that their momentum distributions can overlap. In the overlapping region, the momentum conservation may be readily satisfied; thus recombination of the

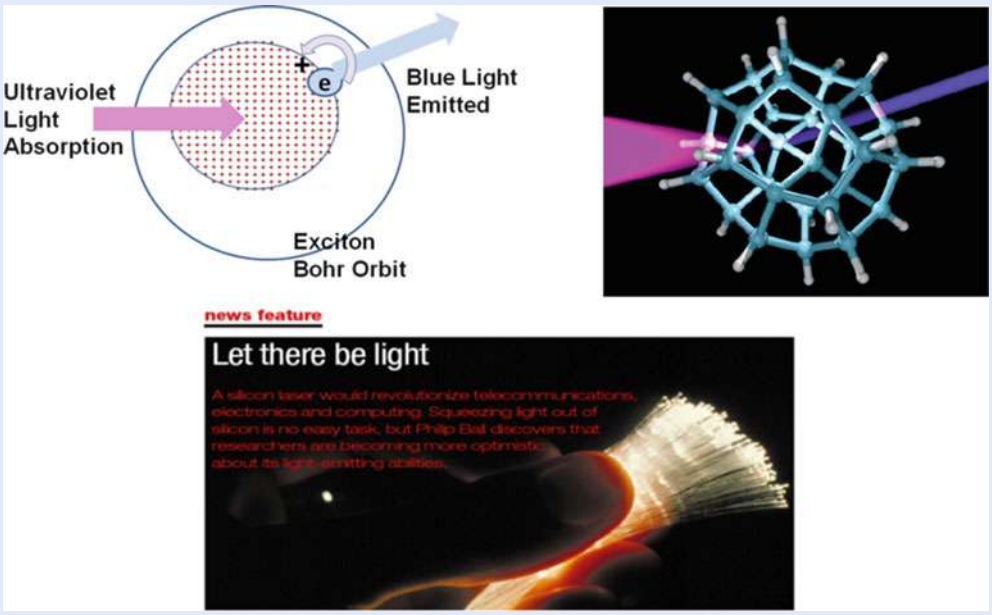


Fig. 10.18 (*Top left*) Schematic of response of 1-nm silicon nanoparticle to light. In 1-nm nanoparticle, much smaller than the Bohr radius, recombination produces a blue photon very fast much before vibrations and heat is produced. (*Top right*) Computer simulation image of 1-nm nanoparticle from [34]. (*Bottom*) Nature magazine report on the brightness of the silicon nanoparticles in terms of a biblical theme "Let there be light" (Image from [38])

electron and hole to turn the excess energy into light becomes stronger and faster than producing vibration and heat. In this case a photon (light) is produced, with the energy of the emitted photon (color of the light emitted) depending on the size of the particle.

For a nanoparticle size of 1 nm, for example, emission of blue photons dominates the production of vibration and heat making it very bright with a performance that matches particles of direct bandgap material as in Fig. 10.18 (top left) [39]. Figure 10.18 (top right) is a computer simulation image of 1-nm nanoparticle carried out by researchers at Livermore National Laboratory [35]. Making silicon the optically dullest material in the universe glow is highly dramatic. Nature magazine reported this effect using the theme "Let there be light," as depicted in Fig. 10.18 (bottom); it is a biblical theme of creation, signaling the transformation of the universe from the dark state to the bright state [38].

10.3.5 Optical Nonlinearity in Nanosilicon

The structure of a silicon crystal has cubic symmetry which exhibits inversion symmetry (centrosymmetry). But when the crystal becomes very small, it loses its rigidity and atoms especially the ones on the surface re-adjust, which exerts strain across the entire particle that alters the cubic structure and breaks the symmetry. Breaking the symmetry produces fundamental effects on its interaction with light. Figure 10.19 (top left) displays one theoretical process that takes place in 1-nm particle in which some of the surface atoms lose their hydrogen atoms, move as much as 1.5 Å each towards each other followed by reconstruction/connection to form Si–Si dimer-like systems.

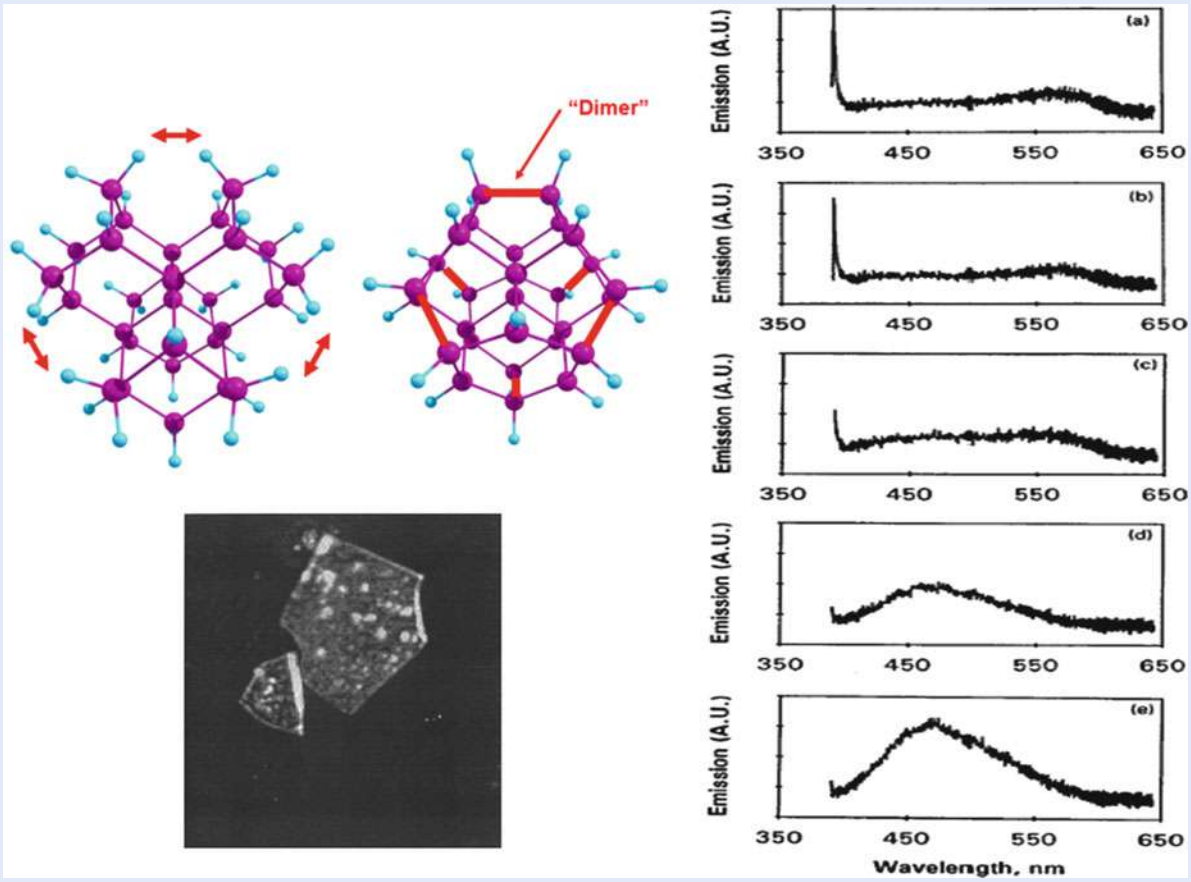


Fig. 10.19 Response of microcrystallites of 1-nm silicon particles to 780-nm near-infrared femtosecond excitation. (*Top left*) Model of surface reconstruction that breaks the crystal symmetry of silicon. (*Left bottom*) Image of reconstituted microcrystals. (*Right*) Shows strong blue luminescent band indicating two-photon excitation as well as competing sharp radiation at half the wavelength of the incident beam. The intensity of the second harmonic anti correlates with the wide band luminescence

10

Blue luminescent 1-nm nanoparticles were used to test this mechanism. Particles prepared in water solvent were dried on device-quality silicon, which allowed them to form microcrystallites films, as shown in Fig. 10.19 (left bottom). The microcrystallites were then excited by high intensity near-infrared femtosecond laser at 800-nm wavelength. Figure 10.19 (right) displays several frames of the emission from different crystallites. It shows strong wide blue luminescent band as well as competing sharp radiation at half the wavelength of the incident beam (second harmonic). The intensity of the second harmonic anticorrelates with the wide band luminescence as seen in the figure. Both the sharp and wide band response indicates very strong coupling with light and the emergence of second-order nonlinear phenomena. The excitation of the blue luminescence band indicates a two photon excitation process, while the second harmonic indicates the breaking of the centrosymmetry. Other theoretical simulations showed that the mechanical strain exerted is extremely high, corresponding to a pressure of several Gpa [40]. Because atoms on the surface are under high strain, they can undergo large-amplitude molecular-like vibrations to relieve the strain [41]; and hence can also couple efficiently to thermal and mechanical stimuli. Second-order nonlinearity in silicon opens up other prospective applications in optics, including modulation, amplification, gain and laser action, and signal processing [42].

10.3.6 Optical Gain in Nanosilicon-Based Material

Not only miniaturization afforded silicon strong luminescence and optical nonlinearity, but there are also some preliminary reports indicating optical gain and laser action in films of silicon nanoparticles [43]. In one report [44], silicon nanoparticles were created inside a glass slab by silicon ion implantation followed by high temperature annealing that allows the silicon atoms to condense into silicon nanoparticles of ~3-nm diameter. The slab acted as an optical amplifier of weak light beams. In other reports 1-nm silicon nanoparticles were reconstituted into microcrystallites [45]. The crystallites were irradiated by infrared femtosecond pulses of a laser beam with high peak power. Microscopic directed blue emission was observed. Microscopic red emission was also observed from clusters of 2.9-nm nanoparticles under excitation of strong incoherent green light [46].

10.4 Applications of Optics in Nanotechnology

Miniaturization has triggered strong light–matter interactions. Nanotechnology allowed researchers to study light–matter interactions at the nanoscale and to launch the subfield of nano-optics. A considerable amount of basic knowledge as well as novel functions of nanomatter has accumulated over the past 20 years. Some applications have already reached the practical level, while others are futuristic. In this section we briefly present some practical applications of optics in nanotechnology in service of fields as diverse as electronics, opto- and photoelectronics, elementary particles, biomedicine, energy harvest and lighting, and art as in stained glass and lusterware pottery.

10.4.1 Integration of Optics and Electronics

Bulk semiconductors, especially silicon, form the backbone of modern electronics and computing. Bulk silicon, however, is a very dull material, being an especially poor emitter of light, turning added energy into heat. This makes integrating electronic and photonic circuits a challenge. There have been several proposals to alleviate this problem. One approach involves doping silicon with other materials. However, the emitted light is not in the visible rather in deep infrared. Moreover, the emission is not very efficient and can degrade the electronic properties of silicon. Another approach is to use nanotechnology by making silicon devices that are very small, such as using luminescent nanoparticles, five nanometers in diameter or less as introduced above. As we have seen above, at that size quantum confinement effects allow the device to emit light. But making electrical connections at that scale is not currently feasible and may compromise the optical activity of the nanomaterial [47], as well as afford very low electrical conductivity.

Another scheme involves subjecting bulk silicon to concentrated fields of plasmonics. The high fields afforded by plasmonic nanometal material can cause distortion and modification of the crystal structure, hence the interaction with light [48–50]. We describe briefly an architecture that demonstrates some of these effects. First a pure silicon nanowire is wrapped with a coating of glass, as shown in Fig. 10.20 (left). Then it is mounted on a glass substrate. Then it is coated with silver. Because of the glass substrate, silver does not wrap completely around the wire, effectively making the silver coating take the shape Ω. This leaves a narrow transparent glass window through which a laser beam can be sent in and emitted light can be extracted [48]. Practically, this narrow window did not compromise

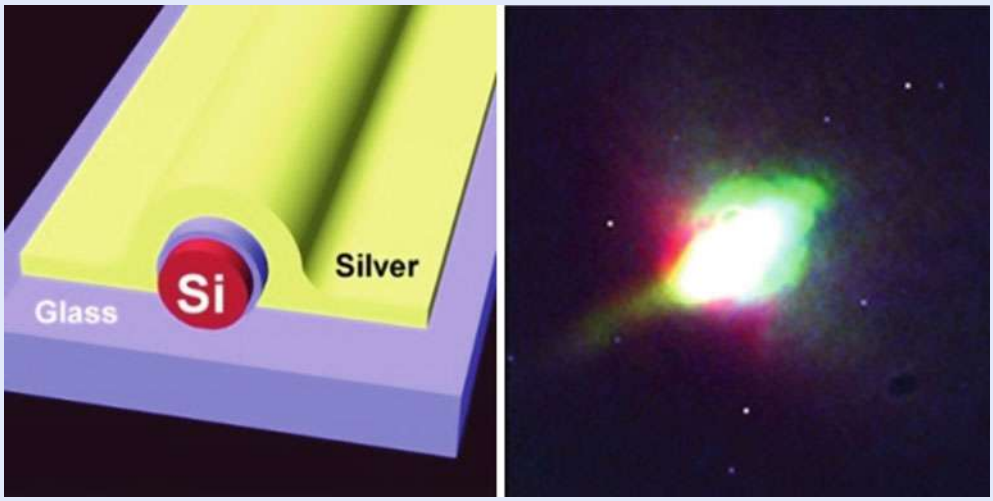


Fig. 10.20 Schematic of a hybrid silicon–silver nanowire system. (*Left*) Silicon nanowire wrapped with a glass coating followed by silver coating in the shape of Ω, on a glass substrate. (*Right*) white light emission when a laser beam irradiates the system through the Ω window. (From [48])

the silver coating from acting as a plasmonic cavity. When a narrow band blue laser beam is sent through the window, the silicon wire produces white light that spans the visible spectrum, as shown in Fig. 10.20 (right). It is to be noted that bare silicon wires do not produce such white light when excited by the blue laser, indicating plasmonic effects. The broad bandwidth light provides good potential for operation in photonic or optoelectronic devices.

10.4.2 Confined Light in Service of Substance Detection

Understanding interactions between strong light and matter is central to many fields. Deeper understanding of strong light coupling with matter is expected to afford the creation of advanced tailored applications [50]. The designation "strong field" applies to an external electromagnetic field that is sufficiently strong to cause significant alterations in atomic or molecular structure and dynamics of the material. Present-day laser technology as well as light enhancement due to plasmonic effects has made it possible to study the behavior of atoms and molecules in fields that have peak electric field strength of the order of atomic fields inside atoms or molecules. Under these conditions, even tightly bound ground states must be greatly altered by the presence of the field [51, 52].

Understanding interactions between light and nanomatter is central to many fields, providing invaluable insights into the nature of matter as well as the nature of light. Indeed, greater understanding of light–matter coupling has enabled creation of tailored applications, resulting in a variety of devices such as microscopic lasers, switches, sensors, modulators, fuel and solar cells, and detectors. As discussed above, hybrid-plasmonic monolithic nanowire optical cavities highlight recent progress made in tailoring light–matter coupling strengths.

Detailed Maxwell's theory was used to calculate the enhancement in intensity of light and the corresponding electric fields of light in the proximity of gold nanoparticles [53]. The results for 10-nm nanoparticles are given in Fig. 10.21a, using a gray/white color code for incident light of 1 V/m electric field. The enhancement based on these theoretical estimations can be as large as a factor of 10, which provides enhancements of two orders of magnitude in intensity.

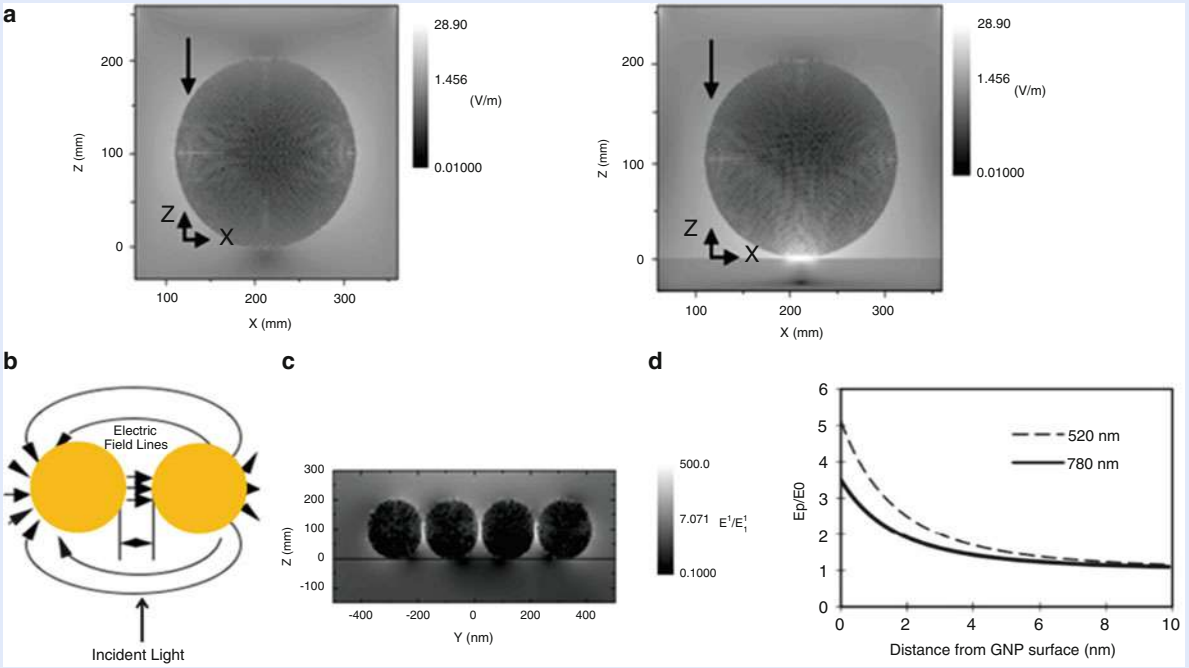
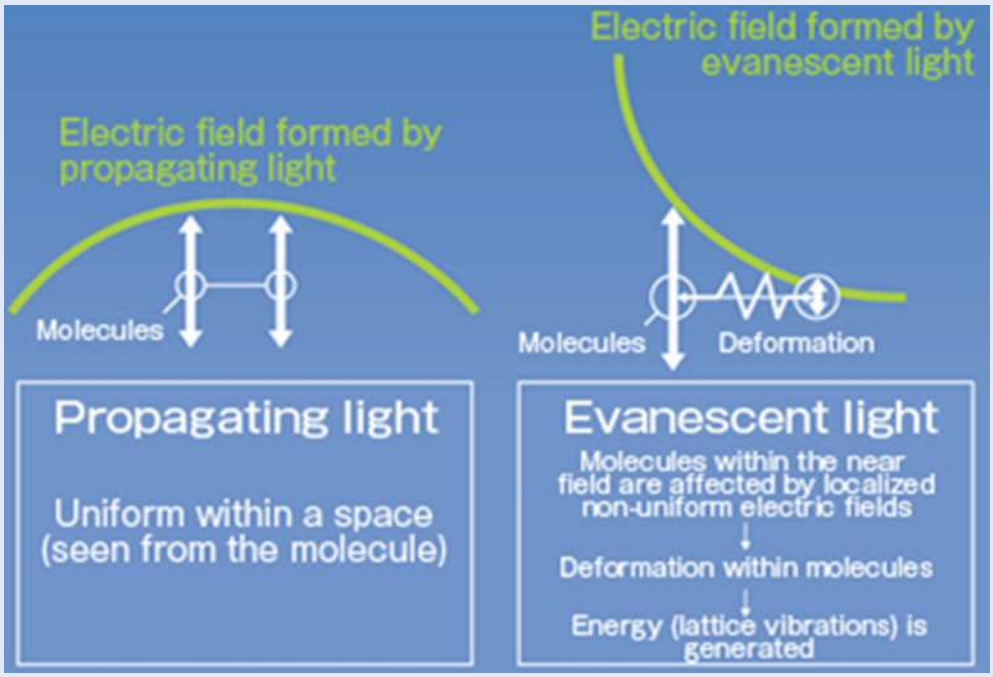


Fig. 10.21 Detailed Maxwell's theory was used to calculate the intensity of light and the corresponding electric fields of light in the proximity of 10-nm gold nanoparticles. (**a**) Near a single nanoparticles (**b**) schematic of the field lines between two particles as well as (**c**) actual results. *Gray/white color code* is used for incident light of 1 V/m electric field (**d**) detailed dependence, as an example, of the electric field near a gold nanoparticle illuminated with light of two different wavelengths. (From [53])

Enhancements can be even more dramatic when there is more than one metal particle. Figure 10.21b, c displays a schematic of the field lines between two particles as well as actual results using the gray/white color code, respectively. It shows, in space between two closely placed plasmonic nanoparticles the electric field may be enhanced by nearly a factor of 30, providing nearly three orders of magnitude enhancement in intensity. Figure 10.21d gives detailed dependence on the wavelength of the light. As an example, the figure gives the electric field near a gold nanoparticle illuminated with light of two different wavelengths [53]

Thus the confined light to nanometal spaces is short range (dies out exponentially) and does not propagate to large distances. Because of the short range of the field, it interacts only with nanostructures next to it, within ~10–30 nm. Thus it provides spatial nanoresolution. This opens up many applications, especially those optical processes or phenomena that depend very sensitively on the magnitude of the electric field, such as fluorescence, Raman scattering, and infrared absorption, resulting in plasmon-enhanced fluorescence, surface-enhanced Raman scattering, and surface-enhanced infrared absorption spectroscopy.

The Raman scattering effect, for example, depends on the fourth power of the field; thus they are enhanced if scattering is performed with the confined light near a metal nanoparticle rather than with ordinary propagating light. The Raman scattering effect is used for sensing and identification of substances. When a substance is irradiated with light, it scatters light called Raman light at a wavelength or frequency slightly shifted from the wavelength of the original irradiated light by an amount equal to the natural frequency of molecules that make up the substance. So, detecting the Raman light and analyzing its spectrum allows the



10

Fig. 10.22 Schematic of a large molecule in the proximity of metal nanoparticle. It shows the profile of the plasmonic electric field. Different parts of the molecule experience different electric field strengths, introducing strong distortion within the molecules (From [16])

identification of the substance. Since Raman light is usually very weak, detecting its intensity directly is quite difficult. If the electrical field strength is increased 10 times by surface plasmon resonance, then the Raman light is intensified 10^4 times so the intensity becomes 10,000 times higher.

Another interesting phenomenon of strong fields is the ability to optically manipulate the structure of molecules. Since the electric field in the vicinity of the plasmonic nanoparticles drops very rapidly over 10 nm, different parts of large molecules will experience different electric field strengths hence different electric stress. This causes distortion within the molecules and consequently causes new ways to deliver energy to the molecule including vibrations, nonlinearities, and electronic structure, as shown in Fig. 10.22. The additional energy supplements the energy of the incident photon [16].

10.4.3 Nanofabrication and Nanolithography

Atoms or molecules are placed near the atomically sharp metal tip (tungsten or gold tip) of a scanning tunneling microscope (STM), as shown in Fig. 10.23 (left top) [54, 55]. Below the tip is a conducting substrate at a distance of 1 nm. When the tip is biased by 1–3 V, an electric field is setup in the gap which can reach more than 100 MV/cm. An intense pulsed laser light bathes the metal tip and the gap and the substrate. The electric field in the vicinity of the gap is very strong due to the dc field of the tip as well as to plasmonic fields. The combined effect of the light and electric field is confined to a nanospace and will be very intense on atoms or molecules falling in this space near the tip. The intense pulsed laser light can be tuned to resonate with a certain state in an atom or molecule, promoting them to high-lying excited states while the combined electric field can assist in stripping the excited electron to produce a single ionization event of a free electron and a free ion.

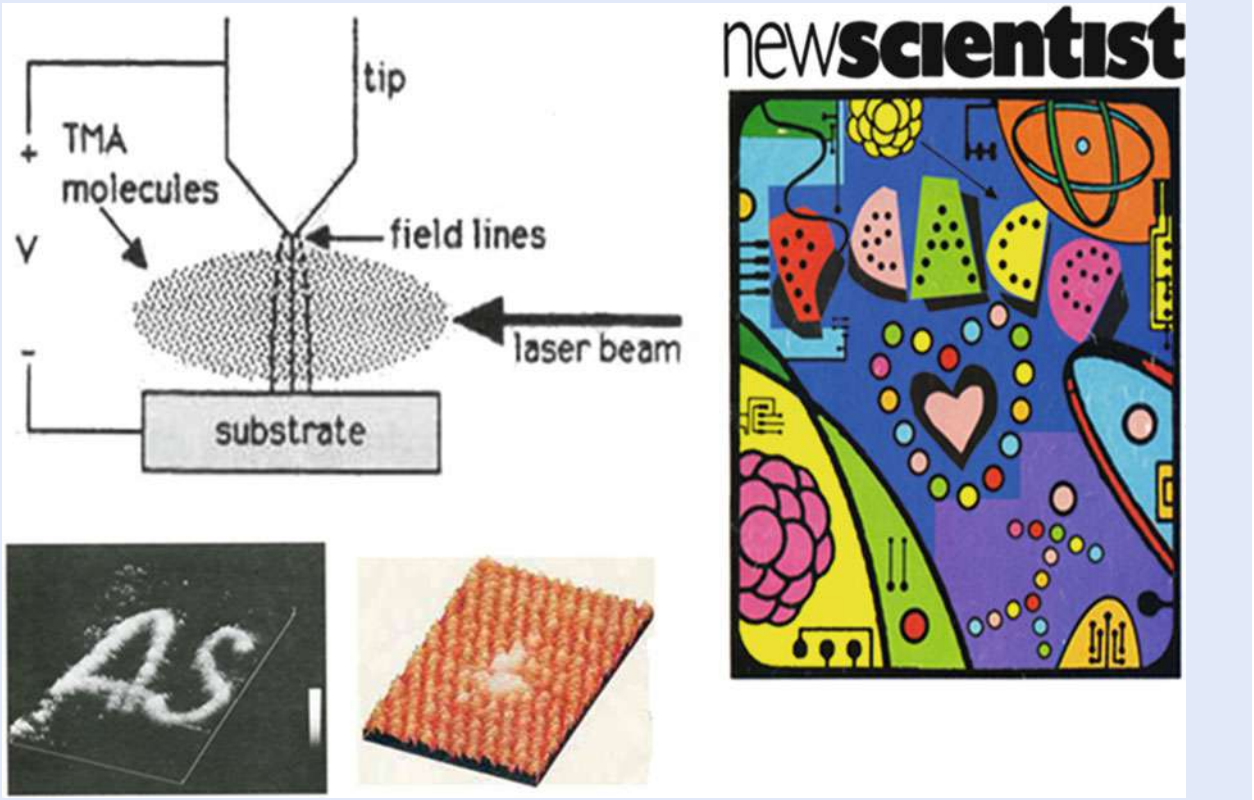


Fig. 10.23 (*Left top*) Schematic of a tungsten tip of a scanning electron microscope bathed with a laser beam in a chamber with a certain pressure of a molecular gas. The metal tip is biased at 1–3 V. Molecules, in the gap or near the gap, are subjected to the combined effect of the laser field and the localized field of the metal tip excites, dissociates, ionizes, and pins down ions on the counter surface. (*Left bottom*) Magnified image of the initials of Arthur Schawlow "*AS*" written with trimethyl aluminum molecules on a silicon wafer. A single molecule of trimethyl aluminum is picked on a graphite surface. (*Right*) Cover of The New Scientist showing displaying a nanofabrication pattern in the form of a "heart" performed by the process depicted in the top left schematic (From [54]) along with patterns in the form of the "molecular man" by IBM and the word "PEACE" by Hitachi, both written using STM tips only

Since the electric field is focused or confined to nanoscale then the atomic ion is accelerated, pushed, and guided to the surface with nanoscale resolution. Programming the position of the tip allowed the Nayfeh's group to make nanopatterns on the surface with atomic resolution. Since light can resonate with specific atoms or molecules then the process is highly selective and can pick a certain atom from a mixture of atoms or molecules, as shown in Fig. 10.23 (left bottom). In one of the earliest news media reports on nanotechnology under the title the "Smallest Graffiti in the World," and subtitle "Nanotechnology Rules," the British magazine, "new Scientist" in 1992 covered the process and displayed several patterns ("heart") on its cover along with the work from IBM ("molecular man"), and Hitachi of Japan ("peace") (Fig. 10.23 (right)) as symbols of nanotechnology [54]. In 1992 on 70th birthday of Arthur Schawlow, the inventor of the laser the author presented him with the nanoscale of his initials "*AS*" written with trimethyl aluminum molecules on a silicon wafer, shown in Fig. 10.23 (left bottom). STM-based processing has been proposed by the community for nanolithography in the electronics industry, but since it relies on the movement of mass for patterning it was concluded that it is very slow and as such unpractical for mass production.

The electric field can also be confined in the vicinity of a very thin tungsten wire, such as that of a Geiger counter used in detecting and counting charged


NEWS

search & discovery

Resonance electron spectroscopy detects single atoms

A group at Oak Ridge National Laboratory reports that by means of the new technique of resonance ionization spectroscopy they have observed single atoms of cesium. They say that one atom in an environment of 10^{19} atoms or molecules of another kind can be measured with their method, and they do not believe this to be a limit. In subsequent experiments, the group has observed density fluctuations of atoms, absolute photodissociation cross sections, time-resolved diffusion and chemical reaction of atoms, and the detection of cesium atoms from the spontaneous fission of californium nuclei.

In 1975 William M. Fairbank Jr (now at Colorado State University), Theodor W. Hänsch and Arthur L. Schawlow (Stanford University) reported detecting as low as 100 atoms/cm^3 using cw resonance fluorescence in sodium vapor. A single atom was contained in the detector beam at any instant. The technique required a steady-state concentration of atoms.

That same year G. Samuel Hurst, Marvin G. Payne, Munir H. Nayfeh, John Judish, E. Bryan Wagner, Chung-hsaun Chen and Jack P. Young (Oak Ridge) developed the resonance ionization spectroscopy method. Each atom in a selected quantum state was converted to a positive ion and a free electron by the absorption of two photons, one of which is resonant with an intermediate state, lying more than half the distance to the ionization continuum. With relatively small energy per pulse in a pulsed dye laser with a linewidth of several angstroms, the intermediate state comes into quasiequilibrium with the ground state. Then other photons from the same laser pulse photoionize the intermediate state.

The method was initially applied to measuring the number of excited atoms created by proton interactions in a given volume. The group also showed that if enough photons/cm^3 are available, all the excited states would be ionized.

This year, Hurst, Nayfeh and Young

continued on page 20

The single-atom detector at Oak Ridge National Laboratory, with experimenters (left to right) G. Samuel Hurst, Jack P. Young and Munir H. Nayfeh. The group reports that with this device they have detected cesium at a concentration as low as one cesium atom in 10^{19} other atoms.

Published by The New York Academy of Sciences

The Sciences

MAY/JUNE 1977 ONE DOLLAR

Fig. 10.24 (*Left*) Report in search and discovery of Physics Today about the use of intense light combined with an electric field to identify and selectively detect single atoms (From [57]). (*Right*) cover of The Sciences displaying Alhazen conducting some laboratory experiments. Inside, the issue reports on the detection of single atoms using the process described in the Fig. 10.24 top left sketch (From [58])

particles for nuclear radiation research [56]. This configuration has been used in 1975–1977 for identifying and counting single atoms by Hurst, Nayfeh, and Young at Oak Ridge National Laboratory [57]. It is to be noted in 1975 William Fairbank Jr., Theodor Hansch, and Arthur Schawlow at Stanford University reported detecting as low as 100 atoms/cm^3 using cw resonance fluorescence in sodium vapor [57]. The Oak Ridge results were covered in the search and discovery news of Physics Today, as shown in Fig. 10.24 (left) [57], and in The Sciences magazine published by the New York Academy of Sciences shown in Fig. 10.24 (right) [58]. The cover of same issue of "The Sciences" shows a sixteenth century painting of Islamic astronomers at work in their observatory. One of the scientists shown is Alhasan Ibn Alhaytham (Alhazen). Because the process constitutes the ultimate sensitivity in analytic detection of matter, namely a single atom, the work was also covered in Britannica Yearbook of Science and the Future (1979), McGraw Hill Yearbook of Science and Technology (1979), The World Book Science Annual (1978), and more.

10.4.4 Photovoltaics and Photocurrent

One hot area involving the interaction of light with nanosemiconductor in the presence of an external electric field is the generation of voltage that can be stored. A thin film of silicon nanoparticles or capsules of silicon nanoparticles, for example, are placed on top of a silicon-based p-n junction (amorphous, polycrystalline, or monocrystalline silicon solar cell). When light strikes the nanoparticles some light is absorbed. As a result, electron hole (e–h) pairs (excitons) are produced in the nanoparticles. If the electron and hole are separated from each other completely before they could recombine to produce light (luminescence) or

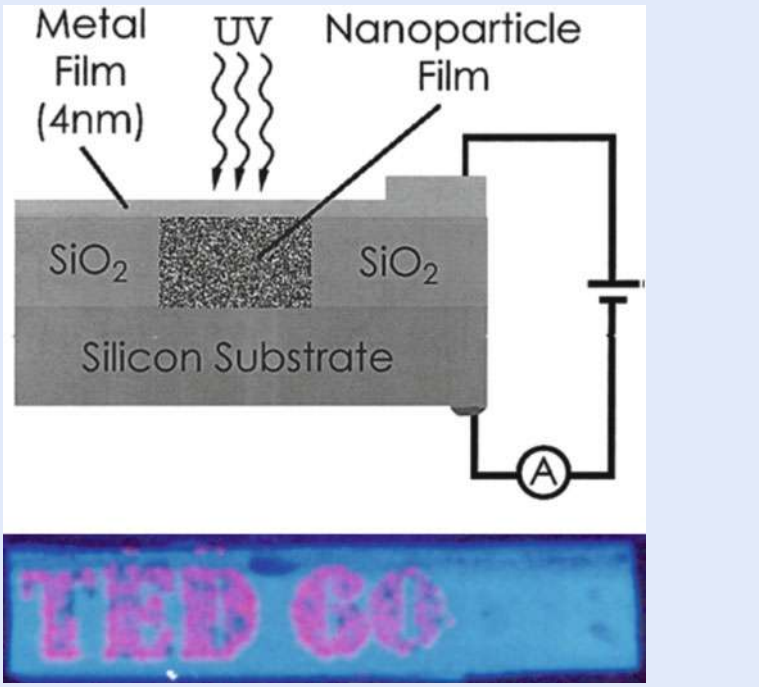


Fig. 10.25 (*Top*) Highly sensitive photodetectors for various applications consisting of holes filled with silicon nanoparticles and covered by an aluminum thin layer (*bottom*) highly magnified image of a pattern of holes in the shape of the words "TED 60" filled with red luminescent silicon nanoparticles under UV excitation was presented to Theodor "Ted" Hansch on his 60th birthday celebration in Munich [62]

recombine non-radiatively to produce vibrations and heat then the electrons and holes may be transported, collected, and stored on two external electrodes appropriately constructed and positioned. The voltage difference between the electrodes can be harnessed at a later time as a battery. The electric field in a p-n junction plays a pivotal role in charge separation and collection. This architecture using 2.9-nm and 1-nm silicon nanoparticles showed relative power enhancements of the efficiency of the under lying solar cell in the UV and in the visible [59, 60, 61].

If the nanoparticles are placed simply on a simple conducting substrate instead of a p-n junction, then a continuous flow of electrons may proceed which may act as an instantaneous current source. A thin oxide layer is grown on a silicon wafer, followed by etching out a hole or multiple holes using a mask. The holes are filled with silicon nanoparticles and an aluminum thin layer is grown as a cap/electrode, as shown in Fig. 10.25 (top). Architectures of this type have been used as highly sensitive photodetectors for various applications [62]. In one application, it provided a sensitive detector for UV radiation while being blind to visible radiation, property which is useful for elementary particle collisions in neutrino, dark matter, and rare decay experiments. In those weak UV Cherenkov radiation is produced that gives vital information about momentum and geometry of collision [63]. Following the same procedure using a mask in the shape of the words "TED 60", a red luminescent pattern was created, as shown in Fig. 10.25 (bottom). In November 2001, the silicon wafer was presented to Theodor "Ted" Hansch on his 60th birthday celebration in Munich. Among the guests are four physics Nobel Laureates: Norman Ramsey, Steven Chu, Claude Cohen-Tannoudgi, and Carl Weiman. The image was published in 2005 in the IEEE Transactions on Nanotechnology just after Hansch received the Nobel Prize [62].

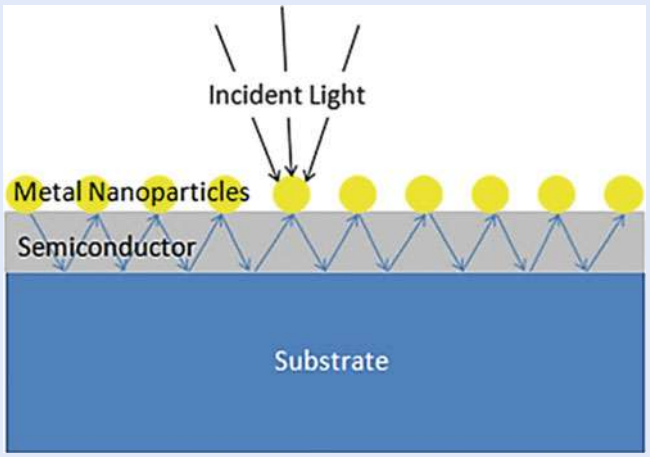


Fig. 10.26 Schematic of a plasmon solar cell. It consists of a thin silicon-based active layer on a glass substrate. Gold nanoparticles are placed on the active layer. (Adapted from ▶ https://en.wikipedia.org/wiki/Plasmonic_solar_cell)

In another development, silicon nanowire arrays were used [64]. Efficient procedures for fabrication of nanowires were recently developed [65, 66]. Hemispherical gold deposits are made on the end of the wires. When a near-infrared optical field falls onto the gold deposits, it excites plasmon resonances, which remarkably amplify the intensity, effectively making them like antennas. The same gold deposit can form a p-n like Schottky junction with the nanowire, which enhances charge collection. Thus the system acts as an effective near-infrared photodetector [64]. Silicon-based sensitive UV photodetectors have military as well as commercial applications. Military applications include missile warning systems, biological attack warning systems, and jet engine sensors.

Light interaction with nanometal is also emerging as a useful intermediate agent for improving light coupling to thin film solar cells; hence improving their efficiency [67]. Thin film solar cells utilize 1–2 μm thick semiconductor material placed over substrates of cheap material compared to silicon, such as glass, plastic, or steel. Metal nanoparticles are deposited on the top of the semiconductor, as shown in Fig. 10.26. When light hits these metal nanoparticles at their surface plasmon resonance, light is scattered in many different directions. This allows light to travel along the active semiconductor and bounce between the substrate and the nanoparticles enabling the semiconductor to absorb more light even though it is a thin layer [67].

10.4.5 Solid State LED White Lighting

The objective of using electronic chips for lighting is to manufacture light bulbs that are superior to conventional bulbs yet with characteristics that match the sun's white light spectrum, given in Fig. 10.27 (top). Light emitting diodes (LED) are chips powered by electricity. They produce light of specific color in the UV, blue, green, or red light range. White light bulbs for domestic use can be manufactured by combing light from three LEDs: blue, green, and red. Because the emission of each LED is not wide enough, the mix produces "finger shaped" spectrum, not smooth as normal sun light (Fig. 10.27 middle). The problem has been somewhat alleviated and simplified in recent years by using a single blue LED source and a wide band green/orange phosphor converter, which combines to

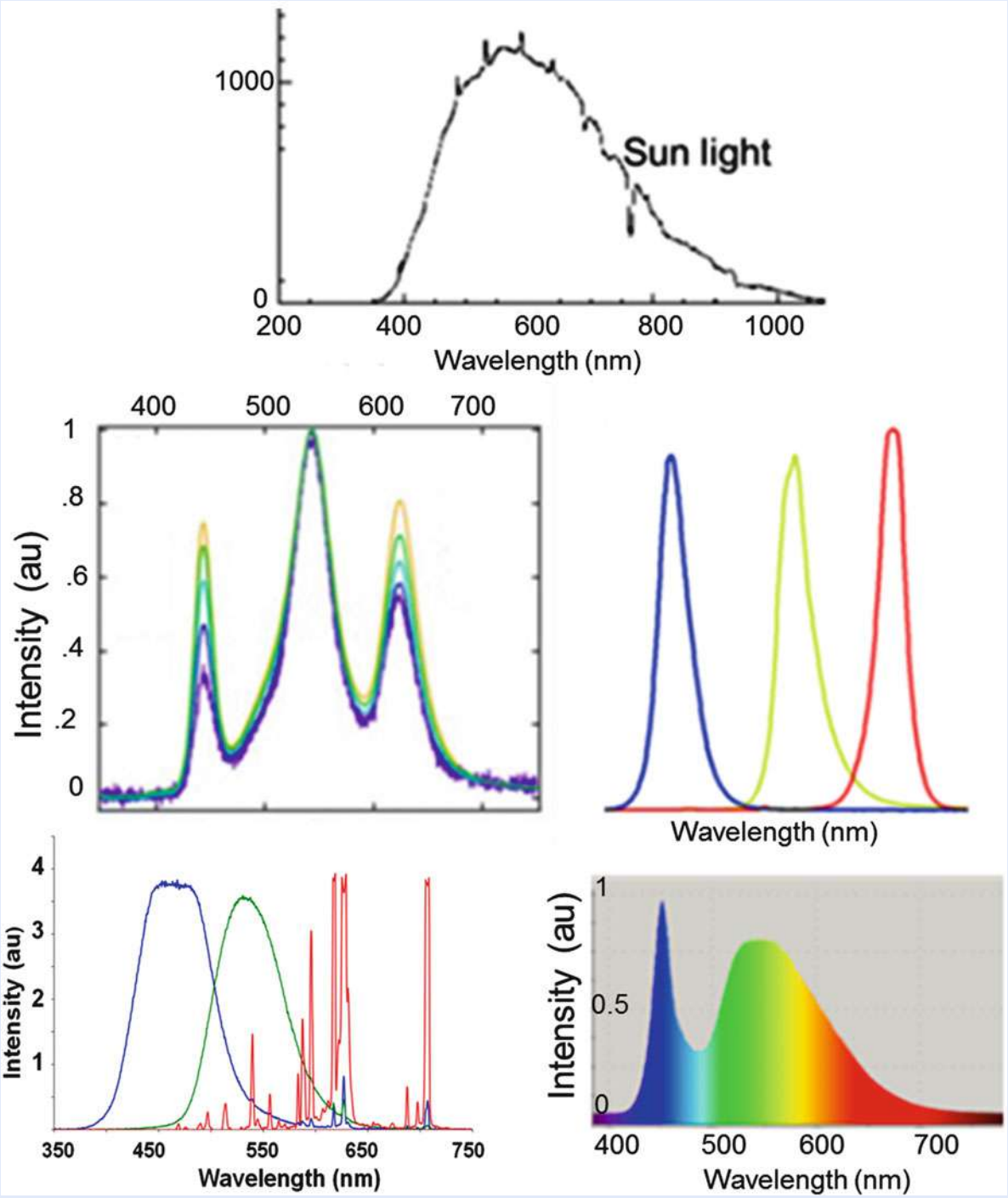


Fig. 10.27 (*Top*) Spectrum of sun light (*middle*) mixing light from three LEDs: *blue*, *green*, and *red*, producing "finger shaped" spectrum not smooth as normal white sun light (*bottom right*) single blue LED source and a green/orange phosphor converter, which combines to a "hand shaped" spectrum not filled enough (*bottom*) single UV LED and three phosphors (blue, green, and red phosphor); red phosphor converters, based on Europium (Eu) provides spectra dominated by *sharp red lines*

a better filled spectrum (Fig. 10.27 bottom right). Although the mix produces a "hand-shaped" spectrum and short of full match with sun light as it is missing a red component, the developers of this process Isamu Akasaki, Hiroshi Amano, and Shuji Nakamura were awarded the 2014 Nobel Prize in Physics. The first LED (red) was invented by Nick Holonyak at the University of Illinois several decades earlier (▶ http://www.led50years.illinois.edu/nicks-story.html), but it was deemed that the more recent blue LED (invented by the awardees) and its use in this configuration afforded a very important service, namely filling the world with new "white" light. A third solution is a UV LED and three wide band phosphor converters (blue, green, and red phosphor). However, there is a problem with red phosphor converters as they are based on europium (Eu). In addition to problems of availability and stability, those provide spectra dominated by sharp red lines (Fig. 10.27 bottom left), which produces a mixed spectrum with sharp lines not smooth enough as normal white sun light (Fig. 10.27 top). In addition, phosphor films are found to have appreciable reflectivity causing a non-negligible fraction of the pumping LED light to be reflected backwards. This in turn causes heating and requires more power input. More energy efficient and rigorous designs must be used to minimize heating.

The novel optical properties of nanomaterial discussed above in Sect. 10.3.3 may alleviate many of these problems. In fact, CdSe-ZnSe quantum dots or nanoparticles have been demonstrated as potential substitute for "red phosphor" for use in near UV pumping [68]. However, due to their direct nature, luminescence of a given nanoparticle size is sharply dependent on its size, which requires the use of an appropriate size distribution to produce broadened emission.

Silicon nanoparticles have also been demonstrated as substitute for or additive to Eu-based "red phosphor" [69], as shown in Fig. 10.28. Because of the indirect nature of silicon, individual silicon nanoparticles produce inhomogeneous broadened luminescence (Fig. 10.28 top), avoiding sharp-line mixing; and because of strong UV absorption they are versatile for pumping with the emerging powerful UV LEDs. It is hoped that UV-pumped "nano-phosphor" converters would allow manufacturing of high quality white bulbs that cover much larger areas while affording high color rendering index which is the quantitative measure of the ability of the bulb to reveal the colors of objects faithfully, while independently providing a correlated color temperature (CCT), which is a specification of the color appearance of the light emitted by the bulb relating its color to the color of light from a black body radiator heated to the same temperature. Commercials CCT labels include warm, daylight, and cool labels of temperature.

Another advantage of using nanomaterial as a component in the phosphor mixture stems from the fact that nanomaterial reduces the reflectivity of the mixed composite material so as not to harm the pumping LED chip. Also nanomaterial improves heat dissipation, thus prolonging the lifetime of the bulb. Thus current manufacturing process of white light can be developed further using the novel interaction of light with nanomatter to produce better efficiency bulbs that cover smoothly the solar white light range, as well as handle larger areas.

10.4.6 Plasmonic Hyperthermic-Based Treatment and Monitoring of Acute Disease

One serious problem inherit in plasmonics is energy loss due to heat generation resulting from strong absorption especially in the infrared—visible and UV spectral ranges [70, 71]. So far, the losses are effectively too high to make practical plasmon-based electronic or photonic devices. However, these losses turned out to be a blessing for other applications, especially for cancer therapy. In this treatment,

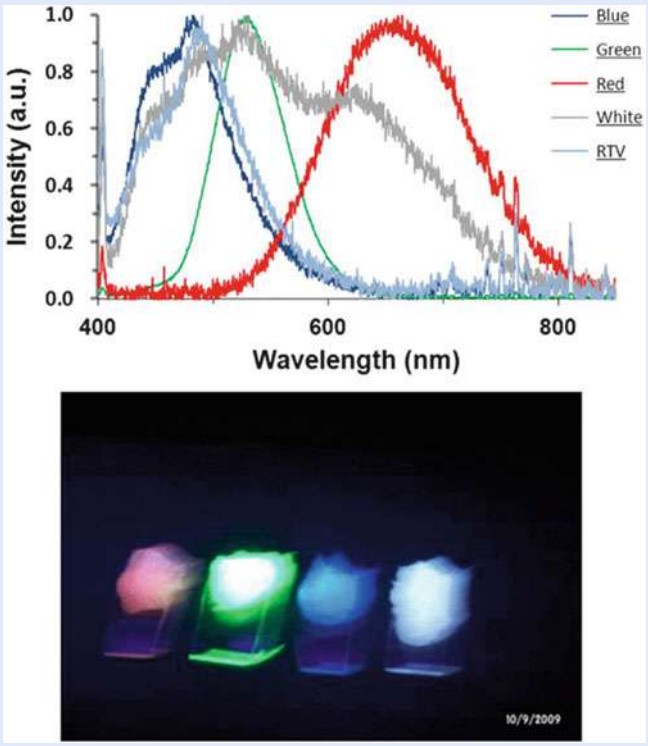


Fig. 10.28 Demonstration of 2.9-nm Si nanoparticles as equivalent to a "red phosphor" for LED technology. (*Top*) Normalized emission spectra of blue phosphor ZnS:Ag (*in blue*), green phosphor ZnS:Cu,Au,Al (*in green*) and red Si nanoparticles (*in red*) individually dispersed in RTV under the excitation of 365 nm radiation. The *gray spectrum* is due to a mixture of the three components dispersed in RTV under excitation of 365 nm. Normalized spectrum of pure RTV (*in light blue*). (*Bottom*) from *left* to *right* the corresponding luminescent images of the individual components and of the mixture under excitation of 365 nm radiation

a near-infrared laser is used, which penetrates deep into the tissue, heating implanted nanoparticle to about 49 °C, as shown in Fig. 10.29 [72]. This is the temperature level which is needed to kill many targeted cancer cells. This results in a threefold increase in killing cancer cells and a substantial tumor reduction within 30 days. Current hypothermic techniques are not selective, i.e., they involve applying heat to the whole body, which heats up cancer cells and healthy tissue, alike. Thus, healthy tissue tends to get damaged. By using gold nanoparticles, which amplify the low energy heat source efficiently, cancer cells can be targeted better and heat damage to healthy tissues can be minimized.

Improvement of the sensitivity of Raman analysis due to the enhanced electric field in the proximity of gold nanoparticles (as discussed above in Sect. 10.4.2.1) has recently been used to monitor and confirm death of cancer cells. The extra sensitivity allows following changes in the molecular content inside cells, including destruction or formation of molecules in cancerous cells during their death [70].

10.5 Plasmon Effect in Ancient Technology and Art

Red glasses have been found in Italy that dated back to the late Bronze Age. The phenomena stayed essentially unsolved till the last decade. It is now believed that the colors are attributed to white light excitation of plasmon surface modes of electron oscillations in different metal nanoparticles incorporated during the

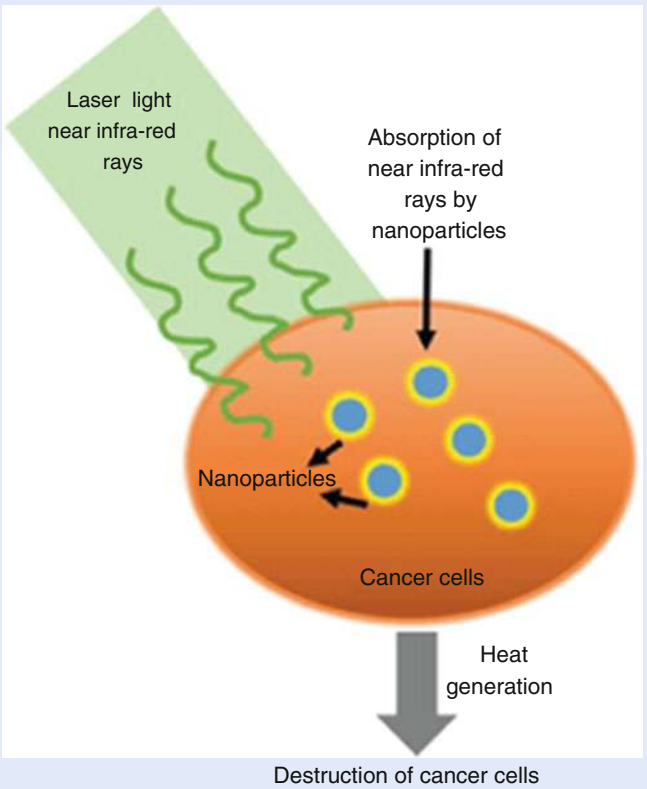


Fig. 10.29 Schematic of plasmonic nanoparticle-based photothermal therapy for cancer treatment. A non-consequential weak infrared light gets concentrated in gold nanoparticles, heating them enough to kill cells (Adapted from [69])

manufacturing process [73, 74]. In addition, gold nanoparticles were identified in some red tesserae of Roman times [75]. A vivid example is the well-known Roman Lycurgus Cup in glass dated from the fourth century CE and currently exhibited in the British Museum, as shown in Fig. 10.30 (left) [76], and the elaborate and complex multicolor stained glass window shown in Fig. 10.30 (right) [77].

As was discussed above in Sect. 10.2.3, the phenomenon can be easily understood in terms of color mixing effects. When white light shines only the G (green) component resonates with the electrons in the gold nanoparticles and is absorbed, hence removed. The B (blue) light is weakened by suffering scattering in all direction. The remaining R (red) component passes through. This is the reason that stained glass mixed with gold nanoparticles appears red to the naked eye.

In the Middle Ages exploitation of the outstanding optical properties of metallic nanoparticles became more sophisticated. Abbasid and Fatimid Islamic potters were able as early as 836–883 CE to develop new chemical technology that allowed them to create complex nanometal-based multilayer structures. First they were incorporated in ceramics, which resulted in the emergence of lusterware, a special type of glazed ceramics, with striking optical effects. Secondly, the distribution of nanoparticles was programmed to include multilayers such that luster decoration color can change depending on the angle from which it is observed. An example of these types of ceramic decorations is given in Fig. 10.31 [78]. The color change is spectacular and brilliant. A variety of very intense colored metallic shine including golden-yellow, blue, green, pink, etc., can be observed. Recent studies in 2005 concluded that the multilayer features give rise to light interference

Fig. 10.30 (*Left*) The well-known Roman Lycurgus Cup in glass dated from the fourth century CE. It is currently exhibited in the British Museum (*Right*) Medieval stained glass windows (Courtesy: NanoBioNet and ► www.nano.gov] [76])

phenomena and scattering through rough interfaces, which adds to the surface plasmon effect and strongly contributes to the observed color [79].

Several material analysis studies [80–83] confirmed that the color of the luster decorations come from metallic nanoparticles. For instance, the presence of metal particles was directly confirmed by conducting high resolution transmission electron microscopy and imaging (TEM) along with material analysis using electron energy loss spectroscopy. Generally, copper and silver clusters were incorporated by applying a mixture of a paint, which contained copper and silver salt powders onto a glazed ceramic. This was followed by annealing in a reducing atmosphere. It is now believed that the basic mechanism involves the two processes of ion exchange and crystallization (nucleation and crystal growth) of copper and silver metallic nanoparticles inside the glassy matrix [84].

The ingenious technology, however, was way ahead of its time hence it was based on empirical chemical means, which made the technology vulnerable to extinction. In fact the technology has nowadays partially been lost. Vincent Reillon and collaborators [85] performed elaborate optical modeling, which confirmed the role of multilayer interference and interface scattering in the color of metallic shine reflection (specular direction). On the experimental side, modern artists [81, 83] conducted experiments to re-create luster decorations. They used modern kilns, which allowed alternative oxidization (oxygen flux) and reduction (CO flux) phases during the firing [73]. The best recreations achieved exhibited partial multilayer structure with a weaker organization.

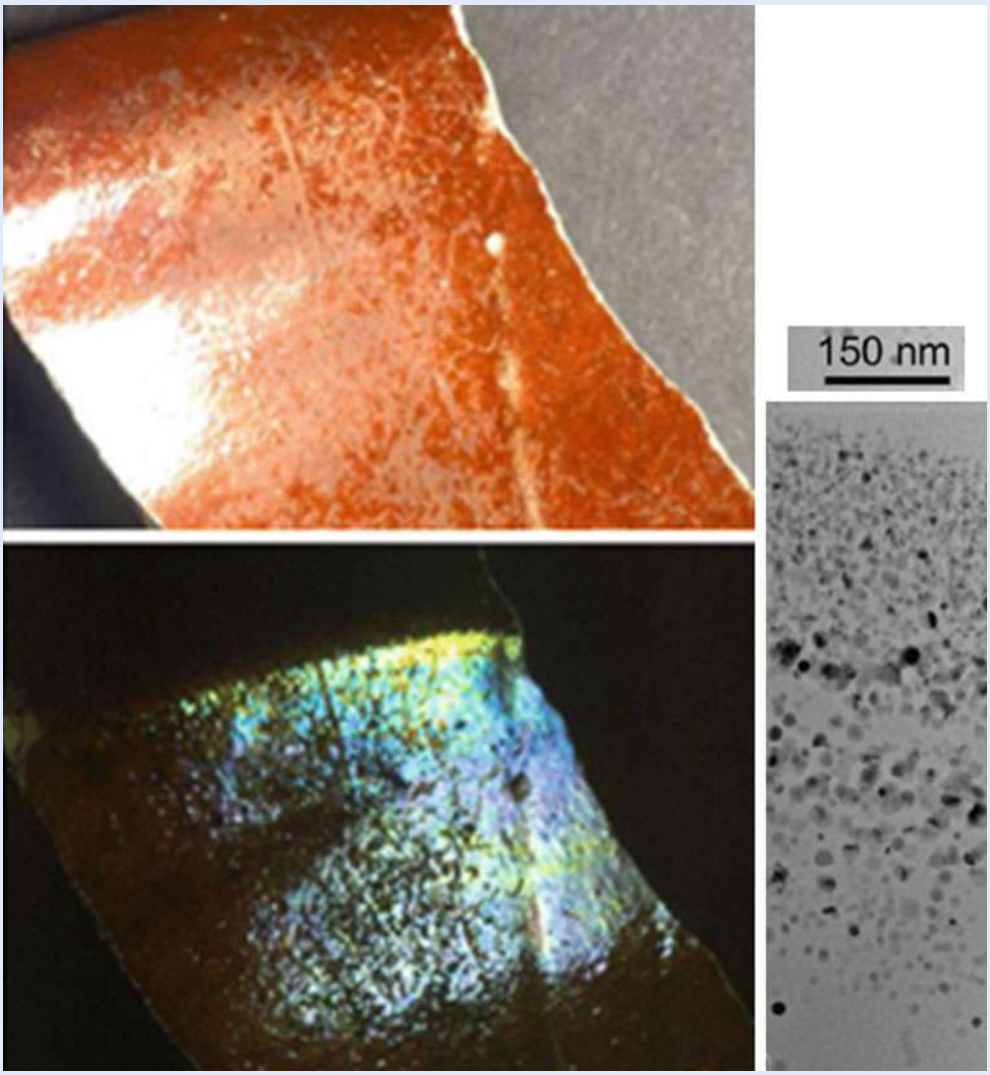


Fig. 10.31 A nineth century CE lusterware from Mesopotamia (Susa). (*Top left*) Lusterware appears *red*. (*Bottom left*) It appears *blue/green* when viewed from different angle. (Micrographs, courtesy D. Chabanne), (*right*) TEM image showing the presence of metal nanoparticle (Adapted from [77, 78])

10.6 Alhasan Ibn Alhaytham (Alhazen) and the Nature of Light and Lusterware

The Abbasid and Fatimid era are considered as two of most creative golden ages of Muslim civilization. The era saw many fascinating advances in science and technology. The period witnessed an aggressive and stimulating advanced chemical and manufacturing technology of glass and pottery, exploiting advanced lighting, color and optical effects of metal nanostructures produced from mixed silver and gold powders. Not only the chemical industry was advanced, material and metallurgical industry including thermo-mechanical forging and annealing was sophisticated. Steel cakes were subjected to such protocols to refine the steel to its exceptional quality. Domestic and military products were produced. One of the products of the industry, the famous Damascus blades, was pivotal in the fight against the Crusaders. Those were forged directly from small cakes of steel (named '*wootz*') imported from ancient India [86, 87]. The blades were remarkably hard,

sharp, and light weight. In fact recent high resolution transmission electron microscopy of samples of the blades revealed the presence of carbon nanotube, which afforded the products exceptional material properties.

Alhasan Ibn Alhaytham (Alhazen) was born in Basra, Iraq in this flourishing era (965). He was not an exception; he contributed to several fields including astronomy, mathematics, medicine, and especially optics. He ventured into pioneering experimental scientific research, with methodology approaching what we use today in the finest research institutes. His studies examined the nature of light and color which resulted in several revolutionary breakthroughs. With limited instrumentation he embarked on thorough examination of the passage of light through various media and its dispersion into its constituent colors, which led him to challenge and refute several adopted notions in optics including:

- Constructing the first generation of pinhole cameras for experimental research
- Establishing, using his pinhole camera, that the human eye is not a source but a sensor or detector of light and color; and accurately described the mechanism of sight and the anatomy of the eye;
- Classifying light sources as luminescent and scattering sources [88, 89];
- Giving the first complete (*unique*) presentation of the law of reflection, by putting the incident ray and the reflected ray in the plane of incidence [90];
- Making the laws of refraction *unique* by introducing the pivotal principle of putting the incident and the refracted rays in the same plane (incidence plane) [90];
- Introducing the reciprocal law of refraction, which Kepler used to discover the law of total reflection [90];
- Proving light travels in straight line and that the speed of light is not infinite rather finite with different speeds in matter, moving more slowly in dense media, which was pivotal to the development of Fermat principle [91];
- Concluding that light is a mix of different colors by conducting experiments on the dispersion of light into its constituent colors; and the discovery that rainbows are caused by refraction of light; those were the first experiments to be conducted before the more detailed experiments conducted by Newton using advances in optical equipment, such as prisms to break up white light into a range of colors [92];
- Studying atmospheric refraction, and using it to discover that twilight only ceases or begins when the sun is 19° below the horizon and to deduce the height of the atmosphere to be 55 miles, which is basically correct (► http://www.jackklaff.com/hos.htm);
- Laying the foundation for the scientific method [93];
- Writing the first comprehensive book on optics, which got translated into Latin, powering research in optics in Europe for many centuries;

These developments show that the quest of Alhazen revolutionized the study of optics at the time and laid the foundation for the scientific method [93]. Alhazen, however, did not try to solve the optical phenomenon of stained glass or lusterware. But those were problems for future generations. For centuries to follow, artists kept mixing silver and gold powder with glass or ceramics to fabricate colorful glass or pottery while no body was able to produce a scientific reason as to how these ingredients worked. The first step towards explaining the phenomenon had to wait until 1908 when Gustav Mie developed a theory of the optical properties of metallic colloids [94]. He showed that the color of a metal nanoparticle depends on its size as well as on the optical properties of the precursor metal and the adjacent dielectric materials. But full understanding of stained glass and lusterware required even more time and nineteenth to twenty-first century developments including electromagnetic theory, electromagnetic wave nature, and interference/diffraction and scattering of light, solid state theory, plasma

theory, nanoscale phenomenon, advanced material characterization technologies, such as electron-based imaging and material analysis. In fact, it was only in the last two decades the scientific community has succeeded in solving the problem [95]. Those medieval artists or chemical industrialists were actually "nanotechnologists" synthesizing metal nanoparticles and harnessing what we today call plasmonics: a new field based on electron oscillations in metals called plasmon.

The recognition of the work of Alhazen may have been slow in the Middle Ages, but it has been coming strong in recent years. In medieval Europe, he was honored as *Ptolemaeus Secundus* ("Ptolemy the Second") or simply called "The Physicist." In the twentieth to twenty-first centuries he has been given several titles including "the First Scientist"; "Hero of Science"; "The Father of Modern Optics"; Pioneering Scientist; Pioneering Scientific Thinker, Rare Genius in Physical Research, The Optical Scientist, etc. In fact, year 2015 marks the 1000th anniversary since the appearance of the Alhazen's remarkable seven volume treatise on optics "Kitab al-Manazir." The year 2015 has also been adopted as the year of light and the United Nations through its arm UNESCO launched "2015 international year of light (IYL2015)" as a global initiative intended to raise awareness of how optical technologies promote sustainable development and provide solutions to worldwide challenges in energy, education, agriculture, communications, and health. One of the major scientific anniversaries that will be celebrated during the 2015 International Year of Light is the works on optics by Alhazen (Ibn Al-Haytham) (1015). UNESCO and the IYL2015 in partnership with the UK-based organization 1001 Inventions launched a high-profile international educational campaign celebrating Ibn al-Haytham. Moreover, 1001 Inventions and the King Abdulaziz Center for World culture in partnership with UNESCO and the IYL2015 are planning to produce a short film on his work. King Abdulaziz City for Science and Technology is also celebrating this occasion by publishing an edited book on optics for a very broad audience, to which this article is contributed.

10.7 From Alhazen to Newton to the Trio: Dispersion of Light

In 1015 Alhazen was concerned with understanding the different color components of light through natural phenomenon, such as propagation of light in material, reflection and refraction, and the rainbow effect. Newton 600 years later used manufactured glass prisms which became available then to deliberately disperse light into its components and to introduce the seven color names red, orange, yellow, green, blue, indigo, and violet for segments of the spectrum. Nearly 300 years after Newton Theodor "Ted" Hansch and Arthur Schawlow of Stanford University and John "Jan" Hall of the University of Colorado set out to achieve big strides in the quest to disperse light and isolate extremely narrow components using much more advanced electronics and optics capabilities, such as prisms, gratings, interference filters, spectrographs, and telescopic light expanders. But such a highly pure color components would have extremely weak intensity, which would necessarily require amplification if it is to be a practical light source. In 1972, the Stanford group dispersed the red fluorescence from a chemical dye into much finer components while simultaneously being amplified using the very advanced light–matter interaction concepts of stimulated emission (introduced by Einstein) and laser gain (Schawlow inventor) to produce a narrowband directed red laser (light amplification by stimulated emission of radiation). Pulses or flashes of red light of 8 ns durations with a record color definition (bandwidth) of 0.0004 Å (or 0.001 cm^{-1} or 30 MHz) and energy of ~1 nJ per pulse were achieved (Hansch laser). In addition to being narrow band, the laser allows for change

(tuning) of wavelength with good control and precision, a pivotal property for matching (resonating with) and studying electronic structure of atoms and molecules [96]. John Hall, on the other hand, used a different approach by starting out with the red light of a helium–neon gas laser [97, 98]. This red laser light at a given time is narrow band (very pure in color) but it is practically not as narrow over longer times because its frequency or wavelength drifts as a result of drift in ambient conditions (temperature, vibrations, etc.). Using advanced optical and electronic equipment he succeeded in making the frequency or wavelength practically stand still for long times. This high finesse light source is extremely useful but it is not tunable in wavelength.

The two groups used these advanced devices in which the color of light has been defined much more precisely to perform high resolution laser spectroscopy measurements of unprecedented accuracy and intrinsic physical interest. Jan Hall used his source to measure accurately the speed of light, allowing a re-definition of the SI meter. It should be mentioned here that the first version of the helium–neon laser was invented in 1960 by Ali Javan (Azerbajani born in Tehran, Iran) at Bell Labs [99]. However the laser produced invisible light at 1.15 μm. Only 18 months later in 1962, Javan's colleagues White and Rigden of Bell Lab constructed the visible He–Ne laser which is more exciting, captivating, and convenient [100].

The Stanford group utilized the tunability of the Hansch laser in a Doppler free technique that was introduced theoretically earlier by Russian scientists [101]. The technique allows the laser to be blind to all moving atoms in a gas sample and see only the ones that happen to be stationary along the direction of the red light beam, which unmasks the true structure of the atoms. Figure 10.32 (left top) is a photo from a celebration at Stanford recognizing Schawlow and Hansch as California Scientist of the year for these developments and the use of the narrowband tunable laser in precise spectroscopic measurements. Issa Shahin, a doctoral student in the Stanford group, was involved in using the narrow bandband dye laser to study the structure of sodium and hydrogen [96]. Munir Nayfeh, a doctoral student under the superposition of Hansch and Schawlow reproduced a variation to the Jan Hall's stable He–Ne light source (Fig. 10.32 left bottom) [103] and utilized it as a wavelength standard along with the narrowband red laser to unmask and measure the hyperfine structure and binding energy of the simplest of all atoms, the hydrogen atom, whose binding energy is the corner stone of all fundamental constants in nature, namely the Rydberg constant. The study achieved then a record accuracy of 9 parts per billion, which allowed the improvements of all other fundamental constants [104]. The importance of improving the accuracy of the Rydberg constant was evident in 1995 when, on the occasion of the 100 year anniversary of the top American Journal Physical Review, the journal listed this achievement as one of the seminal papers in the 100 year life of the Journal, as shown in Fig. 10.32 (right) [102].

Alhazen and Newton and the Trio were concerned with breaking light into its frequency components; other scientists, on the other hand, focused on the time domain. Short pulses of light of duration of micro (10^{-6}), nano (10^{-9}), pico (10^{-12}), femto (10^{-15}), and now atto (10^{-18}) seconds have been isolated. In this regard, we mention that the Egyptian-born Ahmad Zewail used femtosecond pulses of light to capture the very real-time dynamics inside molecules, studies that earned him in 1999 the Nobel Prize in chemistry [105]. In the last decade the interest ventured into the atto second regime, providing capabilities to make real-time observations of valence electron motion in solids [106].

Schawlow, Hansch, and Hall were recognized by Nobel Prizes. Arthur Schawlow received the 1981 Nobel Prize in Physics for his work on development of laser light. Theodor W. Hänsch and John Hall shared with Roy Glauber (for his

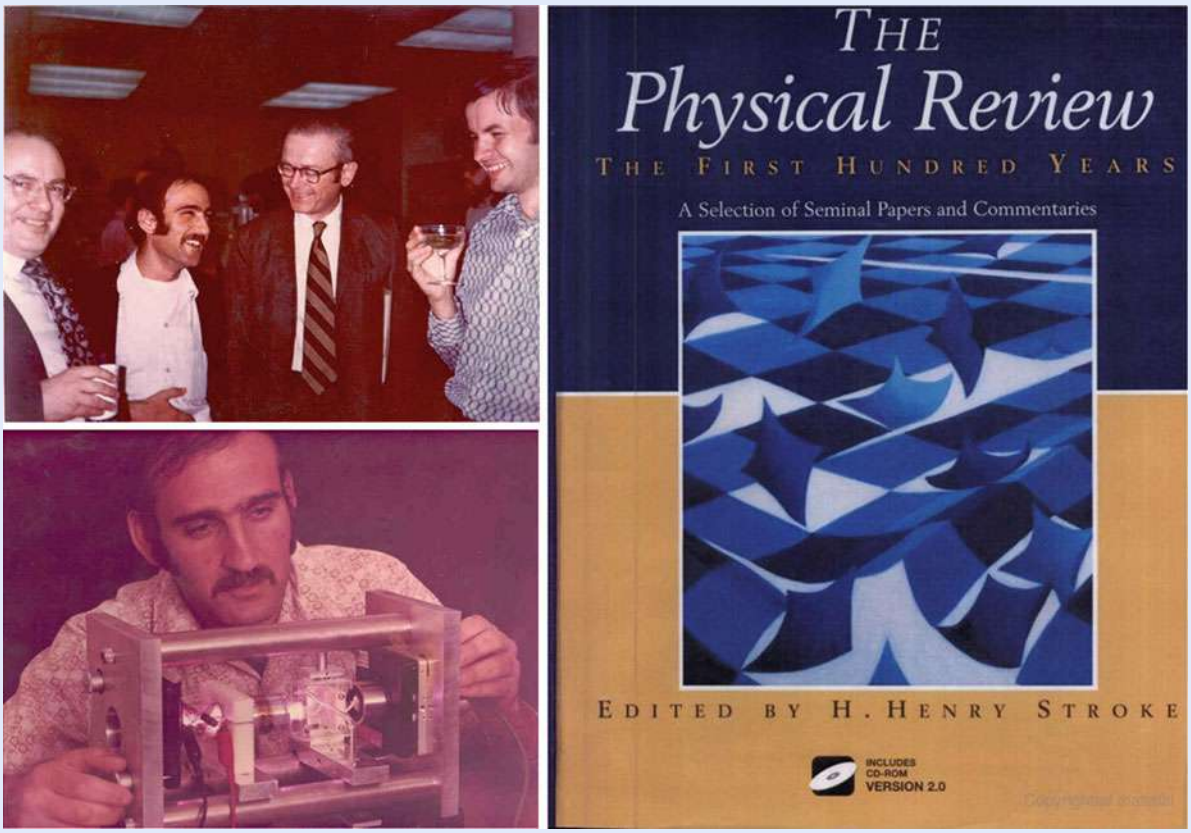


Fig. 10.32 (*Left top*) Celebration at Stanford from left to right: Arthur Schawlow, Munir Nayfeh, Stanford President, and Theodore Hansch for the choice of Schawlow and Hansch as California Scientist of the Year for the high resolution spectroscopy achievements. (*Left bottom*) Home built He–Ne laser by Hansch and Nayfeh after [96] stabilized to 10 parts per billion. It has iodine vapor cell placed in the cavity. It is stabilized by locking it to the *n*th hyperfine component of isotropically pure I^{129} at 633 nm. The photo decorated the office of Arthur Schawlow (*right*) cover of the book celebrating the 100 year anniversary of the American Journal Physical Review [102]

work on coherence of light) the 2005 Nobel Prize in Physics for their work in precision spectroscopy using laser light sources.

It is to be noted that Newton's fame actually comes from gravity (as the father of) rather than from light. Alhazen fame is beginning to come in light, emerging as "the father of modern optics." There were many sources describing Ibn al-Haytham (Alhazen) as such. One of those is "Impact of Science on Society—Volumes 26–27—(1976) Page 140, a prestigious UNESCO publication whose first edition came out in 1950. The study stated that "one name stands out as that of a rare genius in physical research: Abu Ali Al-Hasan Ibn Al-Haytham (965–1039) of Basra (Iraq), without question the father of modern optics" [107].

Figure 10.33 presents selected clips and images from the mass media giving tribute to the contributions of Alhazen. There is one more thing that can be said about Alhazen. If the Nobel Prize was in place in the nineth century or its regulations did not exclude deceased recipients, this author believes Alhazen would have certainly been a recipient for his revolutionary breakthroughs in advancing our understanding of the nature of light.

Fig. 10.33 Selected clips and images from the mass media giving tribute to Alhazen contributions

10.8 Conclusion

We focused in this article on how miniaturization impacts light interaction and propagation in metal and semiconductors. Unlike semiconductors, coupling of light to oscillating free electrons in metals allows metal nanostructures to concentrate or focus light to spots limited only by their size, much beyond the wavelength of incoming light, which enables novel concepts of metal-based lenses. Moreover, strong concentration, absorption, and scattering induce drastic color change as well as temperature rise and interactivity with the surrounding environment. On the other hand, miniaturization of semiconductors affords the material strong size-dependent luminescence and color, to the degree that it can take indirect material, such as silicon from being the dullest material to a glowing material. Integration of plasmonic and semiconductor effects in hybrid architectures is promising synergetic applications.

These novel and unprecedented nanoscale optical functionalities enable a variety of exciting applications including a variety of futuristic applications [108]. But there are novel exciting applications which have been demonstrated and they may be around the corner for the consumer use, such as hyperthermia treatment of cancer, monitoring cell death through molecular changes, substance identification and detection, photovoltaic thin film solar cells, solid state white

lighting, photodetectors, subwavelength waveguides and antennas for commercial, military and elementary particles applications, nanofabrication, integration of optics and electronics, and art of stained glass and lusterware.

Metal plasmonic devices face, however, significant challenges because of heat losses at visible and other high frequencies important for telecommunication. Such heat losses have seriously limited their practical use in electronics; but the heat effect was turned into an asset in the fight on acute disease at the cellular level.

Acknowledgment We acknowledge financial grants that supported our work on Nanoengineered light emitting silicon, including the University of Illinois, US National Science Foundation, US Army, US Office of Naval Research, State of Illinois, Grainger Foundation, Beckman Foundation, King Fahd University, King Saud University and King Abdulaziz City for Science and Technology, Sharp of America, and SunGen of Canada, and Saudi ARAMCO.

■■ Occasion

2015 has been declared as the International Year of Light and Light-based Technologies. This year celebrates many milestones in the history of optics starting from the 1000 year anniversary of Ibn Al-Haytham's achievements in optics among which is his great book on Light. King Abdulaziz City for Science and Technology in Riyadh, Saudi Arabia is celebrating this occasion by publishing an edited book on topics in optics for a very broad audience.

References

1. Editor (1993) McGraw-Hill encyclopedia of science and technology, 5th edn. McGraw-Hill, New York, NY
2. Drexler KE (1986) Engines of creation: the coming era of nanotechnology. Doubleday, New York
3. Feynman R (2009) There's plenty of room at the bottom. Nat Nanotechnol 4:781. doi:► 10.1038/nnano.2009.356. ► http://www.nature.com/nnano/journal/v4/n12/full/nnano.2009.356.html
4. National Nanotechnology Initiative (2016) What is nanotechnology? ► http://www.nano.gov/nanotech-101/what/definition; Benefits and applications, ► http://www.nano.gov/you/nanotechnology-benefits
5. Iijima S, Ichihashi T (1993) Single-shell carbon nanotubes of 1-nm diameter. Nature 363:603–605
6. Nayfeh MH, Rogozhina E, Mitas L (2002) Silicon nanoparticles: next generation of ultrasensitive fluorescent markers. In: Baratron M-I (ed) Synthesis, in functionalization, and surface treatment of nanoparticles. American Scientific Publishers, Stevenson Ranch

7. Nayfeh MH, Mitas L (2007) Silicon nanoparticles: new photonic and electronic material at the transition between solid and molecule. In: Kumar V (ed) Nanosilicon. Elsevier, Amsterdam
8. Lohse SE, Murphy CJ (2013) The quest for shape control: a history of gold nanorod synthesis. Chem Mater 25:1250–1261
9. Babak Nikoobakht B, El-Sayed MA (2003) Preparation and growth mechanism of gold nanorods (NRs) using seed-mediated growth method. Chem Mater 15:1957–1962
10. Murray CB, Kagan CR, Bawendi MG (2000) Synthesis and characterization of monodisperse nanocrystals and close-packed nanocrystal assemblies. Annu Rev Mater Res 30:545–610
11. Foresight Institute (2016) Applications of nanotechnology. ► https://www.foresight.org/nano/applications.html
12. Pines D, Bohm D (1952) A collective description of electron interactions: II. Collective vs individual particle aspects of the interactions. Phys Rev 85:338
13. Messiah A (1999) Quantum mechanics. Dover, Minneola
14. Maier SA (2007) Plasmonics: fundamentals and applications. Springer, New York
15. Ditlbacher H, Hohenau A, Wagner D, Kreibig U, Rogers M, Hofer F, Aussenegg FR, Krenn JR (2005) Silver nano wires as surface plasmon resonators. Phys Rev Lett 95:257403
16. Hamamatsu Photonics KK (2016) Nanophotonics. ► http://www.hamamatsu.com/eu/en/technology/innovation/nanophotonics/index.html
17. Stockman MI (2004) Nanofocusing of optical energy in tapered plasmonic waveguides. Phys Rev Lett 93:137404
18. Evlyukhin AB, Bozhevolnyi SI, Stepanov AL, Kiyan R, Reinhardt C, Passinger S, Chichkov BN (2007) Focusing and directing of surface plasmon polaritons by curved chains of nanoparticles. Opt Express 15:16667–16680
19. Yin L, Vlasko-Vlasov VK, Pearson J, Hiller JM, Hua J, Welp U, Brown DE, Kimball CW (2005) Sub wavelength focusing and guiding of surface plasmons. Nano Lett 5:1399–1402
20. Verslegers L, Catrysse PB, Yu Z, White JS, Barnard ES, Brongersma ML, Fan S (2009) Planar lenses based on nanoscale slit arrays in a metallic film. Nano Lett 9:235–238
21. Veselago VG (1968) The electrodynamics of substances with simultaneously negative values of ε and μ. Sov Phys Usp 10:509–514 (Russian text 1967)
22. West P, Ishii S, Naik G, Emani N, Boltasseva A (2010) Identifying low-loss plasmonic materials, SPIE Newsroom. ► http://www.academia.edu/1037839/Identifying_low-loss_plasmonic_materials
23. White Noise (2014) ► http://whitenoise.kinja.com/dimensions-in-semiconductors-can-something-be-zero-dim-1543310309
24. Bawendi M (2016) ► http://nanocluster.mit.edu/research.php
25. Trwoga PF, Kenyon AJ, Pitt CW (1998) Modeling the contribution of quantum confinement to luminescence from silicon nanoclusters. J Appl Phys 83:3791
26. Canham LT (1990) Silicon quantum wire array fabrication by electro-chemical and chemical dissolution of wafers. Appl Phys Lett 57:1046
27. Heinrich JL, Curtis CL, Credo GM, Kavanagh KL, Sailor MJ (1992) Luminescent colloidal Si suspensions from porous Si. Science 255:66
28. Yamani Z, Thompson H, AbuHassan L, Nayfeh MH (1997) Ideal anodization of silicon. Appl Phys Lett 70:3404
29. Yamani Z, Ashhab S, Nayfeh A, Nayfeh MH (1998) Red to green rainbow photoluminescence from unoxidized silicon nanocrystallites. J Appl Phys 83:3929
30. Ackakir O, Therrien J, Belomoin G, Barry N, Muller J, Gratton E, Nayfeh M (2000) Detection of luminescent single ultrasmall silicon nanoparticle using fluctuation spectroscopy. Appl Phys Lett 76:1857–1859
31. Belomoin G, Therrien J, Smith A, Rao S, Chaieb S, Nayfeh MH (2002) Observation of a magic discrete family of ultrabright Si nanoparticles. Appl Phys Lett 80:841
32. Nielsen D, Abuhassan L, Alchihabi M, Al-Muhanna A, Host J, Nayfeh MH (2007) Currentless anodization of intrinsic silicon powder grains: formation of fluorescent Si nanoparticles. J Appl Phys 101:114302
33. Mitas L, Therrien J, Twesten R, Belomoin G, Nayfeh MH (2001) Effect of surface reconstruction on the structural prototypes of ultrasmall ultrabright Si_{29} nanoparticles. Appl Phys Lett 78:1918
34. Allan G, Delerue C, Lannoo M (1996) Nature of luminescent surface states of semiconductor nanocrystallites. Phys Rev Lett 76:2961
35. Draeger EW, Grossman JC, Williamson AJ, Galli G (2003) Influence of synthesis conditions on the structural and optical properties of passivated silicon nanoclusters. Phys Rev Lett 90:167402
36. Sundholm D (2003) First principles calculations of the absorption spectrum of $Si_{29}H_{36}$. Nano Lett 3:847

37. Lehtonen O, Sundholm D (2005) Density-functional studies of excited states of silicon nanoclusters. Phys Rev B 72:085424
38. Ball P (2001) Let there be light. Nature 409:974
39. Smith A, Yamani Z, Turner J, Habbal S, Granick S, Nayfeh MH (2005) Observation of strong direct-like oscillator strength in the photoluminescence of 1 nm silicon nanoparticles. Phys Rev B 72:205307
40. Mantey K, Zhu A, Boparai J, Nayfeh M, Marsh C, Alchaar G (2012) Observation of linear solid-solid phase transformation in silicon nanoparticles. Phys Rev B 85:085417
41. Rao S, Mantey K, Therrien J, Smith A, Nayfeh M (2007) Molecular behavior in the vibronic and excitonic properties of hydrogenated silicon nanoparticles. Phys Rev B 76:155316
42. Nayfeh M, Akcakir O, Belomoin G, Barry N, Therrien J, Gratton E (2000) Second harmonic generation in microcrystallite films of ultrasmall Si nanoparticles. Appl Phys Lett 77:4086
43. Kanemitsu Y (2003) Luminescence from Si/Si02 nanostructures. In: Pavesi L et al (eds) Towards the first silicon laser. Kluwer Academic, Dordrecht, pp 109–122
44. Pavesi L, Dal Negro L, Mazzoleni C, Franzò G, Priolo F (2000) Optical gain in silicon nanocrystals. Nature 408:440–444
45. Nayfeh MH, Barry N, Therrien J, Akcakir O, Gratton E, Belomoin G (2001) Stimulated blue emission in reconstituted films of ultrasmall silicon nanoparticles. Appl Phys Lett 78:1131
46. Nayfeh MH, Chaieb S, Rao S, Barry N, Therrien J, Belomoin G, Smith A (2002) Observation of laser oscillation in aggregates of ultrasmall silicon nanoparticles. Appl Phys Lett 80:121
47. Hilliard JE, Nayfeh HM, Nayfeh MH (1995) Re-establishment of photoluminescence in Cu quenched porous silicon by acid treatment. J Appl Phys 77:4130
48. Cho C-H, Aspetti CO, Park J, Agarwal R (2013) Silicon coupled with plasmon nanocavities generates bright visible hot luminescence. Nat Photonics 7:285–289
49. Gu Z, Liu S, Sun S, Wang K, Lyu Q, Xiao S, Song Q (2015) Photon hopping and nanowire based hybrid plasmonic waveguide and ring-resonator. Nat Sci Rep 5:917
50. Piccione B, Aspetti CO, Cho C-H, Agrawal R (2014) Tailoring light–matter coupling in semiconductor and hybrid-plasmonic nanowires. Rep Prog Phys 77:086401
51. Nicholaides C, Nayfeh MH, Clark CW (eds) (1989) Atoms in strong fields. Plenum, New York
52. Glab W, Nayfeh MH (1985) Stark induced resonances in the photoionization of hydrogen. Phys Rev A 31:530–532
53. Nedyalkov N, Imamova S, Atanasov P, Obara M (2010) Gold nanoparticles as nanoheaters and nanolenses in the processing of different substrate surfaces. J Phys Conf Ser 223:012035
54. Clery D (1992) Nanotechnology rules, OK! New Sci 1811:42
55. Yau S-T, Saltz D, Nayfeh MH (1990) Laser-assisted deposition of nanometer structures using scanning tunneling microscopy. Appl Phys Lett 57:2913
56. Hurst GS, Nayfeh MH, Young JP (1977) One-atom detection using resonance ionization spectroscopy. Phys Rev A15:2283
57. Lubkin G (1977) Resonance electron spectroscopy detects single atoms. Phys Today 30:17
58. Quanta (1977) One-atom chemistry. The Sciences. The New York Academy of Sciences, Quanta, p 5, May/June1977
59. Stupca M, Alsalhi M, Al Saud T, Almuhanna A, Nayfeh MH (2007) Enhancement of polycrystalline silicon solar cells using ultrathin films of silicon nanoparticle. Appl Phys Lett 91:063107
60. Maximenko Y, Elhalawany N, Yamani Z, Yau S-T, Nayfeh MH (2013) Polyaniline – Si nanoparticles nanocapsules as a dual photovoltaic sensitizer. Mater Res Soc Symp Proc 1500
61. Chowdhury FI, Nayfeh MH, Nayfeh AM (2016) Enhancened performance of thin film silicon solar cells with a top film of silicon nanoparticles due to down-conversion and near resonance charge transport. J Sol Energy 125:332–338
62. Nayfeh M, Rao S, Nayfeh O, Smith A, Therrien J (2005) UV photodetectors with thin film Si nanoparticle active mediaum. IEEE Trans Nanotechnol 4:660
63. Magill S, Xie J, Nayfeh M, Yu H, Fizari M, Malloy J, Maximenko Y (2015) Enhanced UV light detection using wavelength-shifting properties of Silicon nanoparticles. J Instrum 10: P05008
64. Jee S-W, Zhou K, Kim D-W, Lee J-H (2014) A silicon nanowire photodetector using Au plasmonic nanoantennas. Nano Convergence 1:29
65. Qiu T, Wu XL, Mei YF, Wan GJ, Chu PK, Siu GS (2005) From Si nanotubes to nanowires: synthesis, characterization, and self-assembly. J Cryst Growth 277:143

66. Mantey K, Shams S, Nayfeh MH, Nayfeh O, Alhoshan M, Alrokayan S (2010) Synthesis of wire-like silicon nanostructures by dispersion of SOI using electroless etching. J Appl Phys 108:124321
67. Catchpole KR, Polman A (2008) Plasmonic solar cells. Opt Express 16: 21793–21800. ► http://www.opticsinfobase.org/oe/abstract.cfm?URI=oe-16-26-21793
68. Song H, Lee S (2007) Red light emitting solid state hybrid quantum dot–near-UV GaN LED devices. Nanotechnology 18:255202
69. Stupca M, Nayfeh OM, Hoang T, Nayfeh MH, Alhreish B, Boparai J, Aldwayyan A, AlSalhi M (2012) Silicon nanoparticle-ZnS nanophosphors for UV- based white LED. J Appl Phys 112:074313
70. Kirui DK et al (2013) Targeted near-IR hybrid magnetic nanoparticles for in vivo cancer therapy and imaging. Nanomed Nanotechnol Biol Med 9:702–711
71. Kang B, Mackey MA, El-Sayed MA (2010) Nuclear targeting of gold nanoparticles in cancer cells induces DNA damage, causing cytokinesis arrest and apoptosis. J Am Chem Soc 132:1517–1519
72. Chen J, Keltner L, Christophersen J, Zheng F, Krouse M et al (2002) New technology for deep light distribution in tissue for phototherapy. Cancer J 8:154–163
73. Angelini I, Artioli G, Bellintani P, Diella V, Gemmi M, Polla A, Rossi A (2004) Chemical analyses of bronze age glasses from Frattesina di Rovigo, northern Italy. J Archaeol Sci 31:1175–1184
74. Artioli G, Angelini I, Polla A (2008) Crystals and phase transitions in protohistoric glass materials. Phase Transit 81:233–252
75. Colomban P, March G, Mazerolles L, Karmous T, Ayed N, Ennabli A, Slim H (2003) Raman identification of materials used for jewelry and mosaics in Ifriqiya. J Raman Spectrosc 34:205–213
76. Freestone I, Meeks N, Sax M, Higgitt C (2007) The Lycurgus cup - a Roman nanotechnology. Gold Bull 40:270–277
77. Mallmann M (2008) Medieval stained glass window. Science in School. Courtesy: NanoBioNet and, ► www.nano.gov. ► http://www.scienceinschool.org/2008/issue10/nanotechnology
78. Mirguet C, Roucau C, Sciau P (2009) Transmission electron microscopy a powerful means to investigate the glazed coating of ancient ceramics. J Nano Res 8:141–146
79. Chabanne D (2005) Le décor de lustre métallique des céramiques glaçurées (IXème-XVIIème siècles), Matériaux, couleurs et techniques. PhD thesis, University Bordeaux 3
80. Bobin O, Schvoerer M, Miane JL, Fabre JF (2003) Colored metallic shine associated to luster decoration of glazed ceramics: a theoretical analysis of the optical properties. J Non Cryst Solids 332:28–34
81. Colomban P (2009) The use of metal nanoparticles to produce yellow, red and iridescent color, from bronze age to present times in luster pottery and glass: solid state chemistry, spectroscopy and nanostructure. J Nano Res 8:109–132
82. Lafait J, Berthier S, Andraud C, Reillon V, Boulenguez J (2009) Physical colors in cultural heritage: surface plasmons in glass. C R Phys 10:649–659
83. Cizer S (2010) History, technique and art of Luster. Dokuz Eylul University, Narhdere, Izmir
84. Roqué J, Molera J, Cepria G, Vendrell-Saz M, Perez-Arantegui J (2008) Analytical study of the behaviour of some ingredients used in luster ceramic decorations following different recipes. Phase Transit 81:267–282
85. Reillon V, Berthier S (2006) Modelization of the optical and colorimetric properties of lustered ceramics. Appl Phys Mater Sci Process 83:257–265
86. Sanderson S (2006) Materials: carbon nanotubes in an ancient Damascus sabre, sharpest cut from nanotube sword. Nature 444:286
87. Reibold M, Paufler P, Levin AA, Kochmann W, Pätzke N, Meyer DC (2006) Materials: carbon nanotubes in an ancient Damascus sabre. Nature 444:286
88. Simon G (1996) Vision according to Alhazen: Sciences et savoirs aux XVIe et XVIIe siècles. Presses universitaires du Septentrion: 15
89. Le Guet Tully F (2012) Brief history of astronomical optics. ► https://lise.oca.eu/spip.php?rubrique37
90. Mach E (2003) The principles of physical optics: an historical and philosophical treatment. ► https://books.google.com/books?isbn=0486495590
91. Høg E (2008) 650 Years of optics: from Alhazen to Fermat and Rømer, Astrometry and optics during the past 2000 years - arXiv.org. ► https://arxiv.org/pdf/1104.4554. www.astro.ku.dk/~erik/HoegAlhazen.pdf
92. Color Spaces - color phenomena (2010) ► www.color-theory-phenomena.nl/08.00.html
93. Toler P (2012) Alhazen: the first true scientist? Wonders and Marvels. ► http://www.wondersandmarvels.com/2012/08/alhazan-the-first-true-scientist.html

94. Mie G (1908) Beiträge zur optik trüber medien, speziell kolloidaler metallösungen. Ann Phys 25:377–445
95. Wagner H-P, Kaveh-Baghbadorani M (2015) Plasmonics: revolutionizing light-based technologies via electron oscillations in metals. ► http://phys.org/news/2015-06-plasmonics-revolutionizing-light-based-technologies-electron.html
96. Hänsch TW, Shahin IS, Schawlow AL (1971) High-resolution saturation spectroscopy of the sodium D lines with a pulsed tunable dye laser. Phys Rev Lett 27:707
97. Baer T, Kowalski FV, Hall JL (1980) Frequency stabilization of a 0.633-μm He–Ne longitudinal Zeeman laser. Appl Opt 19:3173–3177
98. Schweitzer WG Jr, Kessler EG Jr, Deslattes RD, Layer HB, Whetstone JR (1973) Description, performance, and wavelengths of iodine stabilized lasers. Appl Opt 12:2927
99. Javan A, Herriott D, Bennett W (1961) Population inversion and continuous optical maser oscillation in a gas discharge containing a He-Ne mixture. Phys Rev Lett 6:106
100. White AD, Rigden JD (1962) Continuous gas maser operation in the visible. Proc IRE 50:1697
101. Vasilenko LS, Chebotayev VP, Shishaev AV (1970) JETP Lett 12:113
102. Stroke HH (ed) (1995) The physical review: the first 100 years: a selection of seminal papers and commentaries. Springer/AIP press, New York
103. Nayfeh MH (1974) Precision measurement of the Rydberg by saturated spectroscopy. PhD thesis, Stanford University
104. Hänsch TW, Nayfeh MH, Lee SA, Curry SM, Shahin IS (1976) Precision measurement of the Rydberg constant by laser saturation spectroscopy of the Balmer line in hydrogen and deuterium. Phys Rev Lett 32:1336
105. Zewail AH (1990) The birth of molecules. Sci Am 263:76
106. Reiter F, Graf U, Serebryannikov EE, Schweinberger W, Fiess M, Schultze M, Azzeer AM, Kienberger R, Krausz F, Zheltikov AM, Goulielmakis E (2010) Route to attosecond nonlinear spectroscopy. Phys Rev Lett 105:243902–243904
107. UNESCO (1976) Impact Sci Soc 26–27:140
108. Focus/Feature (2015) Nano-optics gets practical. Nat Nanotechnol 10:11–15